

THE PRINCETON COMPANION TO
Timothy Gowers

EDITOR

Mathematics



Part I

Introduction

I.1 What Is Mathematics About?

It is notoriously hard to give a satisfactory answer to the question, “What is mathematics?” The approach of this book is not to try. Rather than giving a *definition* of mathematics, the intention is to give a good idea of what mathematics is by describing many of its most important concepts, theorems, and applications. Nevertheless, to make sense of all this information it is useful to be able to classify it somehow.

The most obvious way of classifying mathematics is by its subject matter, and that will be the approach of this brief introductory section and the longer section entitled SOME FUNDAMENTAL MATHEMATICAL DEFINITIONS [I.3]. However, it is not the only way, and not even obviously the best way. Another approach is to try to classify the kinds of questions that mathematicians like to think about. This gives a usefully different view of the subject: it often happens that two areas of mathematics that appear very different if you pay attention to their subject matter are much more similar if you look at the kinds of questions that are being asked. The last section of part I, entitled THE GENERAL GOALS OF MATHEMATICAL RESEARCH [I.4], looks at the subject from this point of view. At the end of that article there is a brief discussion of what one might regard as a third classification, not so much of mathematics itself but of the content of a typical article in a mathematics journal. As well as theorems and proofs, such an article will contain definitions, examples, lemmas, formulas, conjectures, and so on. The point of that discussion will be to say what these words mean and why the different kinds of mathematical output are important.

1 Algebra, Geometry, and Analysis

Although any classification of the subject matter of mathematics must immediately be hedged around with qualifications, there is a crude division that undoubtedly works well as a first approximation, namely the division

of mathematics into algebra, geometry, and analysis. So let us begin with this, and then qualify it later.

1.1 Algebra versus Geometry

Most people who have done some high-school mathematics will think of algebra as the sort of mathematics that results when you substitute letters for numbers. Algebra will often be contrasted with arithmetic, which is a more direct study of the numbers themselves. So, for example, the question, “What is 3×7 ?” will be thought of as belonging to arithmetic, while the question, “If $x + y = 10$ and $xy = 21$, then what is the value of the larger of x and y ?” will be regarded as a piece of algebra. This contrast is less apparent in more advanced mathematics for the simple reason that it is very rare for numbers to appear without letters to keep them company.

There is, however, a different contrast, between algebra and *geometry*, which is much more important at an advanced level. The high-school conception of geometry is that it is the study of shapes such as circles, triangles, cubes, and spheres together with concepts such as rotations, reflections, symmetries, and so on. Thus, the objects of geometry, and the processes that they undergo, have a much more visual character than the equations of algebra.

This contrast persists right up to the frontiers of modern mathematical research. Some parts of mathematics involve manipulating symbols according to certain rules: for example, a true equation remains true if you “do the same to both sides.” These parts would typically be thought of as algebraic, whereas other parts are concerned with concepts that can be visualized, and these are typically thought of as geometrical.

However, a distinction like this is never simple. If you look at a typical research paper in geometry, will it be full of pictures? Almost certainly not. In fact, the methods used to solve geometrical problems very often involve a great deal of symbolic manipulation, although good powers of visualization may be needed to find and use

these methods and pictures will typically underlie what is going on. As for algebra, is it “mere” symbolic manipulation? Not at all: very often one solves an algebraic problem by finding a way to visualize it.

As an example of visualizing an algebraic problem, consider how one might justify the rule that if a and b are positive integers then $ab = ba$. It is possible to approach the problem as a pure piece of algebra (perhaps proving it by induction), but the easiest way to convince yourself that it is true is to imagine a rectangular array that consists of a rows with b objects in each row. The total number of objects can be thought of as a lots of b , if you count it row by row, or as b lots of a , if you count it column by column. Therefore, $ab = ba$. Similar justifications can be given for other basic rules such as $a(b + c) = ab + ac$ and $a(bc) = (ab)c$.

In the other direction, it turns out that a good way of solving many geometrical problems is to “convert them into algebra.” The most famous way of doing this is to use Cartesian coordinates. For example, suppose that you want to know what happens if you reflect a circle about a line L through its center, then rotate it through 40° counterclockwise, and then reflect it once more about the same line L . One approach is to visualize the situation as follows.

Imagine that the circle is made of a thin piece of wood. Then instead of reflecting it about the line you can rotate it through 180° about L (using the third dimension). The result will be upside down, but this does not matter if you simply ignore the thickness of the wood. Now if you look up at the circle from below while it is rotated counterclockwise through 40° , what you will see is a circle being rotated *clockwise* through 40° . Therefore, if you then turn it back the right way up, by rotating about L once again, the total effect will have been a clockwise rotation through 40° .

Mathematicians vary widely in their ability and willingness to follow an argument like that one. If you cannot quite visualize it well enough to see that it is definitely correct, then you may prefer an algebraic approach, using the theory of linear algebra and matrices (which will be discussed in more detail in [I.3 §4.2]). To begin with, one thinks of the circle as the set of all pairs of numbers (x, y) such that $x^2 + y^2 \leq 1$. The two transformations, reflection in a line through the center of the circle and rotation through an angle θ , can both be represented by 2×2 matrices, which are arrays of numbers of the form $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$. There is a slightly complicated, but purely algebraic, rule for multiplying matrices together, and it is designed to have the property that if matrix A represents a transformation R (such as a reflection) and

matrix B represents a transformation T , then the product AB represents the transformation that results when you first do T and then R . Therefore, one can solve the problem above by writing down the matrices that correspond to the transformations, multiplying them together, and seeing what transformation corresponds to the product. In this way, the geometrical problem has been converted into algebra and solved algebraically.

Thus, while one can draw a useful distinction between algebra and geometry, one should not imagine that the boundary between the two is sharply defined. In fact, one of the major branches of mathematics is even called ALGEBRAIC GEOMETRY [IV.7]. And as the above examples illustrate, it is often possible to translate a piece of mathematics from algebra into geometry or vice versa. Nevertheless, there is a definite difference between algebraic and geometric *methods of thinking*—one more symbolic and one more pictorial—and this can have a profound influence on the subjects that mathematicians choose to pursue.

1.2 Algebra versus Analysis

The word “analysis,” used to denote a branch of mathematics, is not one that features at high-school level. However, the word “calculus” is much more familiar, and differentiation and integration are good examples of mathematics that would be classified as analysis rather than algebra or geometry. The reason for this is that they involve *limiting processes*. For example, the derivative of a function f at a point x is the limit of the gradients of a sequence of chords of the graph of f , and the area of a shape with a curved boundary is defined to be the limit of the areas of rectilinear regions that fill up more and more of the shape. (These concepts are discussed in much more detail in [I.3 §5].)

Thus, as a first approximation, one might say that a branch of mathematics belongs to analysis if it involves limiting processes, whereas it belongs to algebra if you can get to the answer after just a finite sequence of steps. However, here again the first approximation is so crude as to be misleading, and for a similar reason: if one looks more closely one finds that it is not so much *branches* of mathematics that should be classified into analysis or algebra, but mathematical *techniques*.

Given that we cannot write out infinitely long proofs, how can we hope to prove anything about limiting processes? To answer this, let us look at the justification for the simple statement that the derivative of x^3 is $3x^2$. The usual reasoning is that the gradient of the chord of the line joining the two points (x, x^3) and $((x+h), (x+h)^3)$

is

$$\frac{(x+h)^3 - x^3}{x+h-x},$$

which works out as $3x^2 + 3xh + h^2$. As h “tends to zero,” this gradient “tends to $3x^2$,” so we say that the gradient at x is $3x^2$. But what if we wanted to be a bit more careful? For instance, if x is very large, are we really justified in ignoring the term $3xh$?

To reassure ourselves on this point, we do a small calculation to show that, whatever x is, the error $3xh + h^2$ can be made arbitrarily small, provided only that h is sufficiently small. Here is one way of going about it. Suppose we fix a small positive number ϵ , which represents the error we are prepared to tolerate. Then if $|h| \leq \epsilon/6x$, we know that $|3xh|$ is at most $\epsilon/2$. If in addition we know that $|h| \leq \sqrt{\epsilon/2}$, then we also know that $h^2 \leq \epsilon/2$. So, provided that $|h|$ is smaller than the minimum of the two numbers $\epsilon/6x$ and $\sqrt{\epsilon/2}$, the difference between $3x^2 + 3xh + h^2$ and $3x^2$ will be at most ϵ .

There are two features of the above argument that are typical of analysis. First, although the statement we wished to prove was about a limiting process, and was therefore “infinitary,” the actual *work* that we needed to do to prove it was entirely finite. Second, the nature of that work was to find sufficient conditions for a certain fairly simple inequality (the inequality $|3xh + h^2| \leq \epsilon$) to be true.

Let us illustrate this second feature with another example: a proof that $x^4 - x^2 - 6x + 10$ is positive for every real number x . Here is an “analyst’s argument.” Note first that if $x \leq -1$ then $x^4 \geq x^2$ and $10 - 6x \geq 0$, so the result is certainly true in this case. If $-1 \leq x \leq 1$, then $|x^4 - x^2 - 6x|$ cannot be greater than $x^4 + x^2 + 6|x|$, which is at most 8, so $x^4 - x^2 - 6x \geq -8$, which implies that $x^4 - x^2 - 6x + 10 \geq 2$. If $1 \leq x \leq \frac{3}{2}$, then $x^4 \geq x^2$ and $6x \leq 9$, so $x^4 - x^2 - 6x + 10 \geq 1$. If $\frac{3}{2} \leq x \leq 2$, then $x^2 \geq \frac{9}{4} \geq 2$, so $x^4 - x^2 = x^2(x^2 - 1) \geq 2$. Also, $6x \leq 12$, so $10 - 6x \geq -2$. Therefore, $x^4 - x^2 - 6x + 10 \geq 0$. Finally, if $x \geq 2$, then $x^4 - x^2 = x^2(x^2 - 1) \geq 3x^2 \geq 6x$, from which it follows that $x^4 - x^2 - 6x + 10 \geq 10$.

The above argument is somewhat long, but each step consists in proving a rather simple inequality—this is the sense in which the proof is typical of analysis. Here, for contrast, is an “algebraist’s proof.” One simply points out that $x^4 - x^2 - 6x + 10$ is equal to $(x^2 - 1)^2 + (x - 3)^2$, and is therefore always positive.

This may make it seem as though, given the choice between analysis and algebra, one should go for algebra. After all, the algebraic proof was much shorter, and makes it obvious that the function is always positive.

However, although there were several steps to the analyst’s proof, they were all easy, and the brevity of the algebraic proof is misleading since no clue has been given about how the equivalent expression for $x^4 - x^2 - 6x + 10$ was found. And in fact, the general question of when a polynomial can be written as a sum of squares of other polynomials turns out to be an interesting and difficult one (particularly when the polynomials have more than one variable).

There is also a third, hybrid approach to the problem, which is to use calculus to find the points where $x^4 - x^2 - 6x + 10$ is minimized. The idea would be to calculate the derivative $4x^3 - 2x - 6$ (an algebraic process, justified by an analytic argument), find its roots (algebra), and check that the values of $x^4 - x^2 - 6x + 10$ at the roots of the derivative are positive. However, though the method is a good one for many problems, in this case it is tricky because the cubic $4x^3 - 2x - 6$ does not have integer roots. But one could use an analytic argument to find small intervals inside which the minimum must occur, and that would then reduce the number of cases that had to be considered in the first, purely analytic, argument.

As this example suggests, although analysis often involves limiting processes and algebra usually does not, a more significant distinction is that algebraists like to work with exact formulas and analysts use estimates. Or, to put it even more succinctly, algebraists like equalities and analysts like inequalities.

2 The Main Branches of Mathematics

Now that we have discussed the differences between algebraic, geometrical, and analytical thinking, we are ready for a crude classification of the subject matter of mathematics. We face a potential confusion, because the words “algebra,” “geometry,” and “analysis” refer *both* to specific branches of mathematics *and* to ways of thinking that cut across many different branches. Thus, it makes sense to say (and it is true) that some branches of analysis are more algebraic (or geometrical) than others; similarly, there is no paradox in the fact that algebraic topology is almost entirely algebraic and geometrical in character, even though the objects it studies, topological spaces, are part of analysis. In this section, we shall think primarily in terms of subject matter, but it is important to keep in mind the distinctions of the previous section and be aware that they are in some ways more fundamental. Our descriptions will be very brief: further reading about the main branches of mathematics can be found in parts II and IV, and more specific points are discussed in parts III and V.

2.1 Algebra

The word “algebra,” when it denotes a branch of mathematics, means something more specific than manipulation of symbols and a preference for equalities over inequalities. Algebraists are concerned with number systems, polynomials, and more abstract structures such as groups, fields, vector spaces, and rings (discussed in some detail in SOME FUNDAMENTAL MATHEMATICAL DEFINITIONS [I.3]). Historically, the abstract structures emerged as generalizations from concrete instances. For instance, there are important analogies between the set of all integers and the set of all polynomials with rational (for example) coefficients, which are brought out by the fact that they are both examples of algebraic structures known as *Euclidean domains*. If one has a good understanding of Euclidean domains, one can apply this understanding to integers and polynomials.

This highlights a contrast that appears in many branches of mathematics, namely the distinction between general, abstract statements and particular, concrete ones. One algebraist might be thinking about groups, say, in order to understand a particular rather complicated group of symmetries, while another might be interested in the general theory of groups on the grounds that they are a fundamental class of mathematical objects. The development of abstract algebra from its concrete beginnings is discussed in THE ORIGINS OF MODERN ALGEBRA [II.3].

A supreme example of a theorem of the first kind is THE INSOLUBILITY OF THE QUINTIC [V.24]—the result that there is no formula for the roots of a quintic polynomial in terms of its coefficients. One proves this theorem by analyzing symmetries associated with the roots of a polynomial, and understanding the group that is formed by them. This concrete example of a group (or rather, class of groups, one for each polynomial) played a very important part in the development of the abstract theory of groups.

As for the second kind of theorem, a good example is THE CLASSIFICATION OF FINITE SIMPLE GROUPS [V.8], which describes the basic building blocks out of which any finite group can be built.

Algebraic structures appear throughout mathematics, and there are many applications of algebra to other areas, such as number theory, geometry, and even mathematical physics.

2.2 Number Theory

Number theory is largely concerned with properties of the set of positive integers, and as such has a consid-

erable overlap with algebra. But a simple example that illustrates the difference between a typical question in algebra and a typical question in number theory is provided by the equation $13x - 7y = 1$. An algebraist would simply note that there is a one-parameter family of solutions: if $y = \lambda$ then $x = (1 + 7\lambda)/13$, so the general solution is $(x, y) = ((1 + 7\lambda)/13, \lambda)$. A number theorist would be interested in *integer* solutions, and would therefore work out for which integers λ the number $1 + 7\lambda$ is a multiple of 13. (The answer is that $1 + 7\lambda$ is a multiple of 13 if and only if λ has the form $13m + 11$ for some integer m .) Other topics studied by number theorists are properties of special numbers such as primes.

However, this description does not do full justice to modern number theory, which has developed into a highly sophisticated subject. Most number theorists are not directly trying to solve equations in integers; instead they are trying to understand structures that were originally developed to study such equations but which then took on a life of their own and became objects of study in their own right. In some cases, this process has happened several times, so the phrase “number theory” gives a very misleading picture of what some number theorists do. Nevertheless, even the most abstract parts of the subject can have down-to-earth applications: a notable example is Andrew Wiles’s famous proof of FERMAT’S LAST THEOREM [V.12].

Interestingly, in view of the discussion earlier, number theory has two fairly distinct subbranches, known as ALGEBRAIC NUMBER THEORY [IV.3] and ANALYTIC NUMBER THEORY [IV.4]. As a rough rule of thumb, the study of equations in integers leads to algebraic number theory and the study of prime numbers leads to analytic number theory, but the true picture is of course more complicated.

2.3 Geometry

A central object of study is the *manifold*, which is discussed in [I.3 §6.9]. Manifolds are higher-dimensional generalizations of shapes like the surface of a sphere, which have the property that any small portion of them looks fairly flat but the whole may be curved in complicated ways. Most people who call themselves geometers are studying manifolds in one way or another. As with algebra, some will be interested in particular manifolds and others in the more general theory.

Within the study of manifolds, one can attempt a further classification, according to when two manifolds are regarded as “genuinely distinct.” A topologist regards

two objects as the same if one can be continuously deformed, or “morphed,” into the other; thus, for example, an apple and a pear would count as the same for a topologist. This means that relative distances are not important to topologists, since one can change them by suitable continuous stretches. A *differential* topologist asks for the deformations to be “smooth” (which means “sufficiently differentiable”). This results in a finer classification of manifolds and a different set of problems. At the other, more “geometrical,” end of the spectrum are mathematicians who are much more concerned by the precise nature of the distances between points on a manifold (a concept that would not make sense to a topologist) and in auxiliary structures that one can associate with a manifold. See RIEMANNIAN METRICS [I.3 §6.10] and RICCI FLOW [III.80] for some indication of what the more geometrical side of geometry is like.

2.4 Algebraic Geometry

As its name suggests, algebraic geometry does not have an obvious place in the above classification, so it is easier to discuss it separately. Algebraic geometers also study manifolds, but with the important difference that their manifolds are defined using polynomials. (A simple example of this is the surface of a sphere, which can be defined as the set of all (x, y, z) such that $x^2 + y^2 + z^2 = 1$.) This means that algebraic geometry is algebraic in the sense that it is “all about polynomials” but geometric in the sense that the set of solutions of a polynomial in several variables is a geometric object.

An important part of algebraic geometry is the study of *singularities*. Often the set of solutions to a system of polynomial equations is similar to a manifold, but has a few exceptional, singular points. For example, the equation $x^2 = y^2 + z^2$ defines a (double) cone, which has its vertex at the origin $(0, 0, 0)$. If you look at a small enough neighborhood of a point x on the cone, then, provided x is not $(0, 0, 0)$, the neighborhood will resemble a flat plane. However, if x is $(0, 0, 0)$, then no matter how small the neighborhood is, you will still see the vertex of the cone. Thus, $(0, 0, 0)$ is a singularity. (This means that the cone is not actually a manifold, but a “manifold with a singularity.”)

The interplay between algebra and geometry is part of what gives algebraic geometry its fascination. A further impetus to the subject comes from its connections to other branches of mathematics. There is a particularly close connection with number theory, explained in ARITHMETIC GEOMETRY [IV.6]. More surprisingly, there are important connections between algebraic geom-

etry and mathematical physics. See MIRROR SYMMETRY [IV.14] for an account of some of these.

2.5 Analysis

Analysis comes in many different flavors. A major topic is the study of PARTIAL DIFFERENTIAL EQUATIONS [IV.16]. This began because partial differential equations were found to govern many physical processes, such as motion in a gravitational field, for example. But they arise in purely mathematical contexts as well—particularly in geometry—so partial differential equations give rise to a big branch of mathematics with many subbranches and links to many other areas.

Like algebra, analysis has some abstract structures that are central objects of study, such as BANACH SPACES [III.64], HILBERT SPACES [III.37], C^* -ALGEBRAS [IV.19 §3], and VON NEUMANN ALGEBRAS [IV.19 §2]. These are all infinite-dimensional VECTOR SPACES [I.3 §2.3], and the last two are “algebras,” which means that one can multiply their elements together as well as adding them and multiplying them by scalars. Because these structures are infinite dimensional, studying them involves limiting arguments, which is why they belong to analysis. However, the extra algebraic structure of C^* -algebras and von Neumann algebras means that in those areas substantial use is made of algebraic tools as well. And as the word “space” suggests, geometry also has a very important role.

DYNAMICS [IV.15] is another significant branch of analysis. It is concerned with what happens when you take a simple process and do it over and over again. For example, if you take a complex number z_0 , then let $z_1 = z_0^2 + 2$, and then let $z_2 = z_1^2 + 2$, and so on, then what is the limiting behavior of the sequence z_0, z_1, z_2, \dots ? Does it head off to infinity or stay in some bounded region? The answer turns out to depend in a complicated way on the original number z_0 . The study of *how* it depends on z_0 is a question in dynamics.

Sometimes the process to be repeated is an “infinitesimal” one. For example, if you are told the positions, velocities, and masses of all the planets in the solar system at a particular moment (as well as the mass of the Sun), then there is a simple rule that tells you how the positions and velocities will be different an instant later. Later, the positions and velocities have changed, so the calculation changes; but the basic rule is the same, so one can regard the whole process as applying the same simple infinitesimal process infinitely many times. The correct way to formulate this is by means of partial differential equations and therefore much of dynamics is

concerned with the long-term behavior of solutions to these.

2.6 Logic

The word “logic” is sometimes used as a shorthand for all branches of mathematics that are concerned with fundamental questions about mathematics itself, notably SET THEORY [IV.1], CATEGORY THEORY [III.8], MODEL THEORY [IV.2], and logic in the narrower sense of “rules of deduction.” Among the triumphs of set theory are GÖDEL’S INCOMPLETENESS THEOREMS [V.18] and Paul Cohen’s proof of THE INDEPENDENCE OF THE CONTINUUM HYPOTHESIS [V.21]. Gödel’s theorems in particular had a dramatic effect on philosophical perceptions of mathematics, though now that it is understood that not every mathematical statement has a proof or disproof most mathematicians carry on much as before, since most statements they encounter *do* tend to be decidable. However, set theorists are a different breed. Since Gödel and Cohen, many further statements have been shown to be undecidable, and many new axioms have been proposed that would make them decidable. Thus, decidability is now studied for *mathematical* rather than philosophical reasons.

Category theory is another subject that began as a study of the processes of mathematics and then became a mathematical subject in its own right. It differs from set theory in that its focus is less on mathematical objects themselves than on what is done to those objects—in particular, the maps that transform one to another.

A *model* for a collection of axioms is a mathematical structure for which those axioms, suitably interpreted, are true. For example, any concrete example of a group is a model for the axioms of group theory. Set theorists study models of set-theoretic axioms, and these are essential to the proofs of the famous theorems mentioned above, but the notion of model is more widely applicable and has led to important discoveries in fields well outside set theory.

2.7 Combinatorics

There are various ways in which one can try to define combinatorics. None is satisfactory on its own, but together they give some idea of what the subject is like. A first definition is that combinatorics is about counting things. For example, how many ways are there of filling an $n \times n$ square grid with 0s and 1s if you are allowed at most two 1s in each row and at most two 1s in each col-

umn? Because this problem asks us to count something, it is, in a rather simple sense, combinatorial.

Combinatorics is sometimes called “discrete mathematics” because it is concerned with “discrete” as opposed to “continuous” structures. Roughly speaking, an object is discrete if it consists of points that are isolated from each other and continuous if you can move from one point to another without making sudden jumps. (A good example of a discrete structure is the *integer lattice* \mathbb{Z}^2 , which is the grid consisting of all points in the plane with integer coordinates, and a good example of a continuous one is the surface of a sphere.) There is a close affinity between combinatorics and theoretical computer science (which deals with the quintessentially discrete structure of sequences of 0s and 1s), and combinatorics is sometimes contrasted with analysis, though in fact there are several connections between the two.

A third definition is that combinatorics is concerned with mathematical structures that have “few constraints.” This idea helps to explain why number theory, despite the fact that it studies (among other things) the distinctly discrete set of all positive integers, is not considered a branch of combinatorics.

In order to illustrate this last contrast, here are two somewhat similar problems, both about positive integers.

- (i) Is there a positive integer that can be written in a thousand different ways as a sum of two squares?
- (ii) Let a_1, a_2, a_3, \dots be a sequence of positive integers, and suppose that each a_n lies between n^2 and $(n+1)^2$. Will there always be a positive integer that can be written in a thousand different ways as a sum of two numbers from the sequence?

The first question counts as number theory, since it concerns a very specific sequence—the sequence of squares—and one would expect to use properties of this special set of numbers in order to determine the answer, which turns out to be yes.¹

The second question concerns a far less structured sequence. All we know about a_n is its rough size—it is fairly close to n^2 —but we know nothing about its more detailed properties, such as whether it is a prime, or a

1. Here is a quick hint at a proof. At the beginning of ANALYTIC NUMBER THEORY [IV.4] you will find a condition that tells you precisely which numbers can be written as sums of two squares. From this criterion it follows that “most” numbers cannot. A careful count shows that if N is a large integer, then there are many more expressions of the form $m^2 + n^2$ with both m^2 and n^2 less than N than there are numbers less than $2N$ that can be written as a sum of two squares. Therefore there is a lot of duplication.

perfect cube, or a power of 2, etc. For this reason, the second problem belongs to combinatorics. The answer is not known. If the answer turns out to be yes, then it will show that, in a sense, the number theory in the first problem was an illusion and that all that really mattered was the rough rate of growth of the sequence of squares.

2.8 Theoretical Computer Science

This branch of mathematics is described at considerable length in part IV, so we shall be brief here. Broadly speaking, theoretical computer science is concerned with efficiency of computation, meaning the amounts of various resources, such as time and computer memory, needed to perform given computational tasks. There are mathematical models of computation that allow one to study questions about computational efficiency in great generality without having to worry about precise details of how algorithms are implemented. Thus, theoretical computer science is a genuine branch of pure mathematics: in theory, one could be an excellent theoretical computer scientist and be unable to program a computer. However, it has had many notable applications as well, especially to cryptography (see MATHEMATICS AND CRYPTOGRAPHY [VII.7] for more on this).

2.9 Probability

There are many phenomena, from biology and economics to computer science and physics, that are so complicated that instead of trying to understand them in complete detail one tries to make probabilistic statements instead. For example, if you wish to analyze how a disease is likely to spread, you cannot hope to take account of all the relevant information (such as who will come into contact with whom) but you can build a mathematical model and analyze it. Such models can have unexpectedly interesting behavior with direct practical relevance. For example, it may happen that there is a “critical probability” p with the following property: if the probability of infection after contact of a certain kind is above p then an epidemic may very well result, whereas if it is below p then the disease will almost certainly die out. A dramatic difference in behavior like this is called a *phase transition*. (See PROBABILISTIC MODELS OF CRITICAL PHENOMENA [IV.26] for further discussion.)

Setting up an appropriate mathematical model can be surprisingly difficult. For example, there are physical circumstances where particles travel in what appears to be a completely random manner. Can one make sense of the notion of a random continuous path? It turns out

that one can—the result is the elegant theory of BROWNIAN MOTION [IV.25]—but the proof that one can is highly sophisticated, roughly speaking because the set of all possible paths is so complex.

2.10 Mathematical Physics

The relationship between mathematics and physics has changed profoundly over the centuries. Up to the eighteenth century there was no sharp distinction drawn between mathematics and physics, and many famous mathematicians could also be regarded as physicists, at least some of the time. During the nineteenth century and the beginning of the twentieth century this situation gradually changed, until by the middle of the twentieth century the two disciplines were very separate. And then, toward the end of the twentieth century, mathematicians started to find that ideas that had been discovered by physicists had huge mathematical significance.

There is still a big cultural difference between the two subjects: mathematicians are far more interested in finding rigorous proofs, whereas physicists, who use mathematics as a tool, are usually happy with a convincing argument for the truth of a mathematical statement, even if that argument is not actually a proof. The result is that physicists, operating under less stringent constraints, often discover fascinating mathematical phenomena long before mathematicians do.

Finding rigorous proofs to back up these discoveries is often extremely hard: it is far more than a pedantic exercise in certifying the truth of statements that no physicist seriously doubted. Indeed, it often leads to further mathematical discoveries. The articles VERTEX OPERATOR ALGEBRAS [IV.13], MIRROR SYMMETRY [IV.14], GENERAL RELATIVITY AND THE EINSTEIN EQUATIONS [IV.17], and OPERATOR ALGEBRAS [IV.19] describe some fascinating examples of how mathematics and physics have enriched each other.

1.2 The Language and Grammar of Mathematics

1 Introduction

It is a remarkable phenomenon that children can learn to speak without ever being consciously aware of the sophisticated grammar they are using. Indeed, adults too can live a perfectly satisfactory life without ever thinking about ideas such as parts of speech, subjects, predicates, or subordinate clauses. Both children and

adults can easily recognize ungrammatical sentences, at least if the mistake is not too subtle, and to do this it is not necessary to be able to explain the rules that have been violated. Nevertheless, there is no doubt that one's understanding of language is hugely enhanced by a knowledge of basic grammar, and this understanding is essential for anybody who wants to do more with language than use it unreflectingly as a means to a nonlinguistic end.

The same is true of mathematical language. Up to a point, one can do and speak mathematics without knowing how to classify the different sorts of words one is using, but many of the sentences of advanced mathematics have a complicated structure that is much easier to understand if one knows a few basic terms of mathematical grammar. The object of this section is to explain the most important mathematical "parts of speech," some of which are similar to those of natural languages and others quite different. These are normally taught right at the beginning of a university course in mathematics. Much of *The Companion* can be understood without a precise knowledge of mathematical grammar, but a careful reading of this article will help the reader who wishes to follow some of the later, more advanced parts of the book.

The main reason for using mathematical grammar is that the statements of mathematics are supposed to be completely precise, and it is not possible to achieve complete precision unless the language one uses is free of many of the vaguenesses and ambiguities of ordinary speech. Mathematical sentences can also be highly complex: if the parts that made them up were not clear and simple, then the unclarities would rapidly accumulate and render the sentences unintelligible.

To illustrate the sort of clarity and simplicity that is needed in mathematical discourse, let us consider the famous mathematical sentence "Two plus two equals four" as a sentence of English rather than of mathematics, and try to analyze it grammatically. On the face of it, it contains three nouns ("two," "two," and "four"), a verb ("equals") and a conjunction ("plus"). However, looking more carefully we may begin to notice some oddities. For example, although the word "plus" resembles the word "and," the most obvious example of a conjunction, it does not behave in quite the same way, as is shown by the sentence "Mary and Peter love Paris." The verb in this sentence, "love," is plural, whereas the verb in the previous sentence, "equals," was singular. So the word "plus" seems to take two objects (which happen to be numbers) and produce out of them a new, single object,

while "and" conjoins "Mary" and "Peter" in a looser way, leaving them as distinct people.

Reflecting on the word "and" a bit more, one finds that it has two very different uses. One, as above, is to link two nouns, whereas the other is to join two whole sentences together, as in "Mary likes Paris and Peter likes New York." If we want the basics of our language to be absolutely clear, then it will be important to be aware of this distinction. (When mathematicians are at their most formal, they simply outlaw the noun-linking use of "and"—a sentence such as "3 and 5 are prime numbers" is then paraphrased as "3 is a prime number and 5 is a prime number.")

This is but one of many similar questions: anybody who has tried to classify all words into the standard eight parts of speech will know that the classification is hopelessly inadequate. What, for example, is the role of the word "six" in the sentence "This section has six subsections"? Unlike "two" and "four" earlier, it is certainly not a noun. Since it modifies the noun "subsection" it would traditionally be classified as an adjective, but it does not behave like most adjectives: the sentences "My car is not very fast" and "Look at that tall building" are perfectly grammatical, whereas the sentences "My car is not very six" and "Look at that six building" are not just nonsense but ungrammatical nonsense. So do we classify adjectives further into numerical adjectives and nonnumerical adjectives? Perhaps we do, but then our troubles will be only just beginning. For example, what about possessive adjectives such as "my" and "your"? In general, the more one tries to refine the classification of English words, the more one realizes how many different grammatical roles there are.

2 Four Basic Concepts

Another word that famously has three quite distinct meanings is "is." The three meanings are illustrated in the following three sentences.

- (1) 5 is the square root of 25.
- (2) 5 is less than 10.
- (3) 5 is a prime number.

In the first of these sentences, "is" could be replaced by "equals": it says that two objects, 5 and the square root of 25, are in fact one and the same object, just as it does in the English sentence "London is the capital of the United Kingdom." In the second sentence, "is" plays a completely different role. The words "less than 10" form an adjectival phrase, specifying a property that numbers may or may not have, and "is" in this sentence is like "is"

in the English sentence “Grass is green.” As for the third sentence, the word “is” there means “is an example of,” as it does in the English sentence “Mercury is a planet.”

These differences are reflected in the fact that the sentences cease to resemble each other when they are written in a more symbolic way. An obvious way to write (1) is $5 = \sqrt{25}$. As for (2), it would usually be written $5 < 10$, where the symbol “ $<$ ” means “is less than.” The third sentence would normally not be written symbolically because the concept of a prime number is not quite basic enough to have universally recognized symbols associated with it. However, it is sometimes useful to do so, and then one must invent a suitable symbol. One way to do it would be to adopt the convention that if n is a positive integer, then $P(n)$ stands for the sentence “ n is prime.” Another way, which does not hide the word “is,” is to use the language of sets.

2.1 Sets

Broadly speaking, a *set* is a collection of objects, and in mathematical discourse these objects are mathematical ones such as numbers, points in space, or even other sets. If we wish to rewrite sentence (3) symbolically, another way to do it is to define P to be the collection, or set, of all prime numbers. Then (3) can be rewritten, “5 belongs to the set P .” This notion of belonging to a set is sufficiently basic to deserve its own symbol, and the symbol used is “ \in .” So a fully symbolic way of writing the sentence is $5 \in P$.

The members of a set are usually called its *elements*, and the symbol “ \in ” is usually read “is an element of.” So the “is” of sentence (3) is more like “ \in ” than “ $=$.” Although one cannot directly substitute the phrase “is an element of” for “is,” one can do so if one is prepared to modify the rest of the sentence a little.

There are three common ways to denote a specific set. One is to list its elements inside curly brackets: $\{2, 3, 5, 7, 11, 13, 17, 19\}$, for example, is the set whose elements are the eight numbers 2, 3, 5, 7, 11, 13, 17, and 19. The majority of sets considered by mathematicians are too large for this to be feasible—indeed, they are often infinite—so a second way to denote sets is to use dots to imply a list that is too long to write down: for example, the expressions $\{1, 2, 3, \dots, 100\}$ and $\{2, 4, 6, 8, \dots\}$ can be used to represent the set of all positive integers up to 100 and the set of all positive even numbers, respectively. A third way, and the way that is most important, is to define a set via a *property*: an example that shows how this is done is the expression $\{x : x \text{ is prime and } x < 20\}$. To read an expression such

as this, one first reads the opening curly bracket as “The set of.” Next, one reads the symbol that occurs before the colon. The colon itself one reads as “such that.” Finally, one reads what comes after the colon, which is the property that determines the elements of the set. In this instance, we end up saying, “The set of x such that x is prime and x is less than 20,” which is in fact equal to the set $\{2, 3, 5, 7, 11, 13, 17, 19\}$ considered earlier.

Many sentences of mathematics can be rewritten in set-theoretic terms. For example, sentence (2) earlier could be written as $5 \in \{n : n < 10\}$. Often there is no point in doing this (as here, where it is much easier to write $5 < 10$) but there are circumstances where it becomes extremely convenient. For example, one of the great advances in mathematics was the use of Cartesian coordinates to translate geometry into algebra and the way this was done was to define geometrical objects as sets of points, where points were themselves defined as pairs or triples of numbers. So, for example, the set $\{(x, y) : x^2 + y^2 = 1\}$ is (or represents) a circle of radius 1 with its center at the origin $(0, 0)$. That is because, by the Pythagorean theorem, the distance from $(0, 0)$ to (x, y) is $\sqrt{x^2 + y^2}$, so the sentence “ $x^2 + y^2 = 1$ ” can be reexpressed geometrically as “the distance from $(0, 0)$ to (x, y) is 1.” If all we ever cared about was which points were in the circle, then we could make do with sentences such as “ $x^2 + y^2 = 1$,” but in geometry one often wants to consider the entire circle as a single object (rather than as a multiplicity of points, or as a property that points might have), and then set-theoretic language is indispensable.

A second circumstance where it is usually hard to do without sets is when one is defining new mathematical objects. Very often such an object is a set together with a *mathematical structure* imposed on it, which takes the form of certain relationships among the elements of the set. For examples of this use of set-theoretic language, see sections 1 and 2, on number systems and algebraic structures, respectively, in SOME FUNDAMENTAL MATHEMATICAL DEFINITIONS [I.3].

Sets are also very useful if one is trying to do *meta-mathematics*, that is, to prove statements not about mathematical objects but about the process of mathematical reasoning itself. For this it helps a lot if one can devise a very simple language—with a small vocabulary and an uncomplicated grammar—into which it is in principle possible to translate all mathematical arguments. Sets allow one to reduce greatly the number of parts of speech that one needs, turning almost all of them into nouns. For example, with the help of the membership

symbol “ \in ” one can do without adjectives, as the translation of “5 is a prime number” (where “prime” functions as an adjective) into “ $5 \in P$ ” has already suggested.¹ This is of course an artificial process—imagine replacing “roses are red” by “roses belong to the set R ”—but in this context it is not important for the formal language to be natural and easy to understand.

2.2 Functions

Let us now switch attention from the word “is” to some other parts of the sentences (1)–(3), focusing first on the phrase “the square root of” in sentence (1). If we wish to think about this phrase grammatically, then we should analyze what sort of role it plays in a sentence, and the analysis is simple: in virtually any mathematical sentence where the phrase appears, it is followed by the name of a number. If the number is n , then this produces the slightly longer phrase, “the square root of n ,” which is a noun phrase that denotes a number and plays the same grammatical role as a number (at least when the number is used as a noun rather than as an adjective). For instance, replacing “5” by “the square root of 25” in the sentence “5 is less than 7” yields a new sentence, “The square root of 25 is less than 7,” that is still grammatically correct (and true).

One of the most basic activities of mathematics is to take a mathematical object and transform it into another one, sometimes of the same kind and sometimes not. “The square root of” transforms numbers into numbers, as do “four plus,” “two times,” “the cosine of,” and “the logarithm of.” A nonnumerical example is “the center of gravity of,” which transforms geometrical shapes (provided they are not too exotic or complicated to have a center of gravity) into points—meaning that if S stands for a shape, then “the center of gravity of S ” stands for a point. A *function* is, roughly speaking, a mathematical transformation of this kind.

It is not easy to make this definition more precise. To ask, “What is a function?” is to suggest that the answer should be a *thing* of some sort, but functions seem to be more like processes. Moreover, when they appear in mathematical sentences they do not behave like nouns. (They are more like prepositions, though with a definite difference that will be discussed in the next subsection.) One might therefore think it inappropriate to ask what kind of object “the square root of” is. Should one not simply be satisfied with the grammatical analysis already given?

As it happens, no. Over and over again, throughout mathematics, it is useful to think of a mathematical phenomenon, which may be complex and very un-thinglike, as a single object. We have already seen a simple example: a collection of infinitely many points in the plane or space is sometimes better thought of as a single geometrical shape. Why should one wish to do this for functions? Here are two reasons. First, it is convenient to be able to say something like, “The derivative of \sin is \cos ,” or to speak in general terms about some functions being differentiable and others not. More generally, functions can have *properties*, and in order to discuss those properties one needs to think of functions as things. Second, many algebraic structures are most naturally thought of as sets of functions. (See, for example, the discussion of groups and symmetry in [I.3 §2.1]. See also HILBERT SPACES [III.37], FUNCTION SPACES [III.29], and VECTOR SPACES [I.3 §2.3].)

If f is a function, then the notation $f(x) = y$ means that f turns the object x into the object y . Once one starts to speak formally about functions, it becomes important to specify exactly which objects are to be subjected to the transformation in question, and what sort of objects they can be transformed into. One of the main reasons for this is that it makes it possible to discuss another notion that is central to mathematics, that of *inverting* a function. (See [I.4 §1] for a discussion of why it is central.) Roughly speaking, the inverse of a function is another function that undoes it, and that it undoes; for example, the function that takes a number n to $n - 4$ is the inverse of the function that takes n to $n + 4$, since if you add four and then subtract four, or vice versa, you get the number you started with.

Here is a function f that cannot be inverted. It takes each number and replaces it by the nearest multiple of 100, rounding up if the number ends in 50. Thus, $f(113) = 100$, $f(3879) = 3900$, and $f(1050) = 1100$. It is clear that there is no way of undoing this process with a function g . For example, in order to undo the effect of f on the number 113 we would need $g(100)$ to equal 113. But the same argument applies to every number that is at least as big as 50 and smaller than 150, and $g(100)$ cannot be more than one number at once.

Now let us consider the function that doubles a number. Can this be inverted? Yes it can, one might say: just divide the number by two again. And much of the time this would be a perfectly sensible response, but not, for example, if it was clear from the context that the numbers being talked about were positive integers. Then one might be focusing on the difference between even and

1. For another discussion of adjectives see ARITHMETIC GEOMETRY [IV.6 §3.1].

odd numbers, and this difference could be encapsulated by saying that odd numbers are precisely those numbers n for which the equation $2x = n$ does *not* have a solution. (Notice that one can undo the doubling process by halving. The problem here is that the relationship is not symmetrical: there is no function that can be undone by doubling, since you could never get back to an odd number.)

To specify a function, therefore, one must be careful to specify two sets as well: the *domain*, which is the set of objects to be transformed, and the *range*, which is the set of objects they are allowed to be transformed into. A function f from a set A to a set B is a rule that specifies, for each element x of A , an element $y = f(x)$ of B . (Not every element of the range needs to be used: consider once again the example of “two times” when the domain and range are both the set of all positive integers.)

The following symbolic notation is used. The expression $f : A \rightarrow B$ means that f is a function with domain A and range B . If we then write $f(x) = y$, we know that x must be an element of A and y must be an element of B . Another way of writing $f(x) = y$ that is sometimes more convenient is $f : x \mapsto y$. (The bar on the arrow is to distinguish it from the arrow in $f : A \rightarrow B$, which has a very different meaning.)

If we want to undo the effect of a function $f : A \rightarrow B$, then we can, as long as we avoid the problem that occurred with the approximating function discussed earlier. That is, we can do it if $f(x)$ and $f(x')$ are different whenever x and x' are different elements of A . If this condition holds, then f is called an *injection*. On the other hand, if we want to find a function g that is undone by f , then we can do so as long as we avoid the problem of the integer-doubling function. That is, we can do it if every element y of B is equal to $f(x)$ for some element x of A (so that we have the option of setting $g(y) = x$). If this condition holds, then f is called a *surjection*. If f is both an injection and a surjection, then f is called a *bijection*. Bijections are precisely the functions that have inverses.

It is important to realize that not all functions have tidy definitions. Here, for example, is the specification of a function from the positive integers to the positive integers: $f(n) = n$ if n is a prime number, $f(n) = k$ if n is of the form 2^k for an integer k greater than 1, and $f(n) = 13$ for all other positive integers n . This function has an unpleasant, arbitrary definition but it is nevertheless a perfectly legitimate function. Indeed, “most” functions, though not most functions that one actually uses, are so arbitrary that they cannot be defined. (Such functions may not be useful as individual objects, but they

are needed so that the set of all functions from one set to another has an interesting mathematical structure.)

2.3 Relations

Let us now think about the grammar of the phrase “less than” in sentence (2). As with “the square root of,” it must always be followed by a mathematical object (in this case a number again). Once we have done this we obtain a phrase such as “less than n ,” which is importantly different from “the square root of n ” because it behaves like an adjective rather than a noun, and refers to a property rather than an object. This is just how prepositions behave in English: look, for example, at the word “under” in the sentence “The cat is under the table.”

At a slightly higher level of formality, mathematicians like to avoid too many parts of speech, as we have already seen for adjectives. So there is no symbol for “less than”: instead, it is combined with the previous word “is” to make the phrase “is less than,” which is denoted by the symbol “ $<$.” The grammatical rules for this symbol are once again simple. To use “ $<$ ” in a sentence, one should precede it by a noun and follow it by a noun. For the resulting grammatically correct sentence to make sense, the nouns should refer to numbers (or perhaps to more general objects that can be put in order). A mathematical “object” that behaves like this is called a *relation*, though it might be more accurate to call it a potential relationship. “Equals” and “is an element of” are two other examples of relations.

As with functions, it is important, when specifying a relation, to be careful about which objects are to be related. Usually a relation comes with a set A of objects that may or may not be related to each other. For example, the relation “ $<$ ” might be defined on the set of all positive integers, or alternatively on the set of all real numbers; strictly speaking these are different relations. Sometimes relations are defined with reference to two sets A and B . For example, if the relation is “ \in ,” then A might be the set of all positive integers and B the set of all sets of positive integers.

There are many situations in mathematics where one wishes to regard different objects as “essentially the same,” and to help us make this idea precise there is a very important class of relations known as *equivalence relations*. Here are two examples. First, in elementary geometry one sometimes cares about shapes but not about sizes. Two shapes are said to be *similar* if one can be transformed into the other by a combination of reflections, rotations, translations, and enlargements (see figure 1); the relation “is similar to” is an

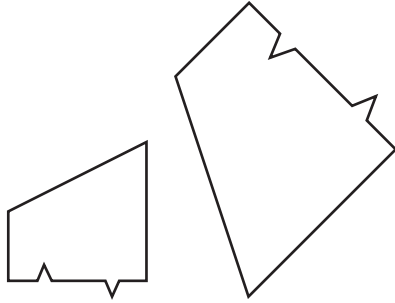


Figure 1 Similar shapes.

equivalence relation. Second, when doing ARITHMETIC MODULO m [III.61], one does not wish to distinguish between two whole numbers that differ by a multiple of m : in this case one says that the numbers are *congruent* (mod m); the relation “is congruent (mod m) to” is another equivalence relation.

What exactly is it that these two relations have in common? The answer is that they both take a set (in the first case the set of all geometrical shapes, and in the second the set of all whole numbers) and split it into parts, called *equivalence classes*, where each part consists of objects that one wishes to regard as essentially the same. In the first example, a typical equivalence class is the set of all shapes that are similar to some given shape; in the second, it is the set of all integers that leave a given remainder when you divide by m (for example, if $m = 7$ then one of the equivalence classes is the set $\{\dots, -16, -9, -2, 5, 12, 19, \dots\}$).

An alternative definition of what it means for a relation \sim , defined on a set A , to be an equivalence relation is that it has the following three properties. First, it is *reflexive*, which means that $x \sim x$ for every x in A . Second, it is *symmetric*, which means that if x and y are elements of A and $x \sim y$, then it must also be the case that $y \sim x$. Third, it is *transitive*, meaning that if x , y , and z are elements of A such that $x \sim y$ and $y \sim z$, then it must be the case that $x \sim z$. (To get a feel for these properties, it may help if you satisfy yourself that the relations “is similar to” and “is congruent (mod m) to” both have all three properties, while the relation “ $<$,” defined on the positive integers, is transitive but neither reflexive nor symmetric.)

One of the main uses of equivalence relations is to make precise the notion of QUOTIENT [I.3 §3.3] constructions.

2.4 Binary Operations

Let us return to one of our earlier examples, the sentence “Two plus two equals four.” We have analyzed the word “equals” as a relation, an expression that sits between the noun phrases “two plus two” and “four” and makes a sentence out of them. But what about “plus”? That also sits between two nouns. However, the result, “two plus two,” is not a sentence but a noun phrase. That pattern is characteristic of *binary operations*. Some familiar examples of binary operations are “plus,” “minus,” “times,” “divided by,” and “raised to the power.”

As with functions, it is customary, and convenient, to be careful about the set to which a binary operation is applied. From a more formal point of view, a binary operation on a set A is a function that takes pairs of elements of A and produces further elements of A from them. To be more formal still, it is a function with the set of all pairs (x, y) of elements of A as its domain and with A as its range. This way of looking at it is not reflected in the notation, however, since the symbol for the operation comes between x and y rather than before them: we write $x + y$ rather than $+(x, y)$.

There are four properties that a binary operation may have that are very useful if one wants to manipulate sentences in which it appears. Let us use the symbol $*$ to denote an arbitrary binary operation on some set A . The operation $*$ is said to be *commutative* if $x * y$ is always equal to $y * x$, and *associative* if $x * (y * z)$ is always equal to $(x * y) * z$. For example, the operations “plus” and “times” are commutative and associative, whereas “minus,” “divided by,” and “raised to the power” are neither (for instance, $9 - (5 - 3) = 7$ while $(9 - 5) - 3 = 1$). These last two operations raise another issue: unless the set A is chosen carefully, they may not always be defined. For example, if one restricts one’s attention to the positive integers, then the expression $3 - 5$ has no meaning. There are two conventions one could imagine adopting in response to this. One might decide not to insist that a binary operation should be defined for every pair of elements of A , and to regard it as a desirable extra property of an operation if it *is* defined everywhere. But the convention actually in force is that binary operations *do* have to be defined everywhere, so that “minus,” though a perfectly good binary operation on the set of all integers, is not a binary operation on the set of all positive integers.

An element e of A is called an *identity* for $*$ if $e * x = x * e = x$ for every element x of A . The two most obvious examples are 0 and 1, which are identities for “plus” and “times,” respectively. Finally, if $*$ has an identity e and

x belongs to A , then an *inverse* for x is an element y such that $x * y = y * x = e$. For example, if $*$ is “plus” then the inverse of x is $-x$, while if $*$ is “times” then the inverse is $1/x$.

These basic properties of binary operations are fundamental to the structures of abstract algebra. See FOUR IMPORTANT ALGEBRAIC STRUCTURES [I.3 §2] for further details.

3 Some Elementary Logic

3.1 Logical Connectives

A *logical connective* is the mathematical equivalent of a conjunction. That is, it is a word (or symbol) that joins two sentences to produce a new one. We have already discussed an example, namely “and” in its sentence-linking meaning, which is sometimes written by the symbol “ \wedge ,” particularly in more formal or abstract mathematical discourse. If P and Q are statements (note here the mathematical habit of representing not just numbers but any objects whatsoever by single letters), then $P \wedge Q$ is the statement that is true if and only if both P and Q are true.

Another connective is the word “or,” a word that has a more specific meaning for mathematicians than it has for normal speakers of the English language. The mathematical use is illustrated by the tiresome joke of responding, “Yes please,” to a question such as, “Would you like your coffee with or without sugar?” The symbol for “or,” if one wishes to use a symbol, is “ \vee ,” and the statement $P \vee Q$ is true if and only if P is true or Q is true. This is taken to include the case when they are both true, so “or,” for mathematicians, is always the so-called *inclusive* version of the word.

A third important connective is “implies,” which is usually written “ \Rightarrow .” The statement $P \Rightarrow Q$ means, roughly speaking, that Q is a consequence of P , and is sometimes read as “if P then Q .” However, as with “or,” this does not mean quite what it would in English. To get a feel for the difference, consider the following even more extreme example of mathematical pedantry. At the supper table, my young daughter once said, “Put your hand up if you are a girl.” One of my sons, to tease her, put his hand up on the grounds that, since she had not added, “and keep it down if you are a boy,” his doing so was compatible with her command.

Something like this attitude is taken by mathematicians to the word “implies,” or to sentences containing the word “if.” The statement $P \Rightarrow Q$ is considered to be true under all circumstances except one: it is not true if P is true and Q is false. This is the *definition* of “implies.” It

can be confusing because in English the word “implies” suggests some sort of connection between P and Q , that P in some way causes Q or is at least relevant to it. If P causes Q then certainly P cannot be true without Q being true, but all a mathematician cares about is this logical consequence and not whether there is any reason for it. Thus, if you want to prove that $P \Rightarrow Q$, all you have to do is rule out the possibility that P could be true and Q false at the same time. To give an example: if n is a positive integer, then the statement “ n is a perfect square with final digit 7” implies the statement “ n is a prime number,” not because there is any connection between the two but because no perfect square ends in a 7. Of course, implications of this kind are less interesting mathematically than more genuine-seeming ones, but the reward for accepting them is that, once again, one avoids being confused by some of the ambiguities and subtle nuances of ordinary language.

3.2 Quantifiers

Yet another ambiguity in the English language is exploited by the following old joke that suggests that our priorities need to be radically rethought.

- (4) Nothing is better than lifelong happiness.
- (5) But a cheese sandwich is better than nothing.
- (6) Therefore, a cheese sandwich is better than lifelong happiness.

Let us try to be precise about how this play on words works (a good way to ruin any joke, but not a tragedy in this case). It hinges on the word “nothing,” which is used in two different ways. The first sentence means “There is no single thing that is better than lifelong happiness,” whereas the second means “It is better to have a cheese sandwich than to have nothing at all.” In other words, in the second sentence, “nothing” stands for what one might call the null option, the option of having nothing, whereas in the first it does not (to have nothing is not better than to have lifelong happiness).

Words like “all,” “some,” “any,” “every,” and “nothing” are called *quantifiers*, and in the English language they are highly prone to this kind of ambiguity. Mathematicians therefore make do with just two quantifiers, and the rules for their use are much stricter. They tend to come at the beginning of sentences, and can be read as “for all” (or “for every”) and “there exists” (or “for some”). A rewriting of sentence (4) that renders it unambiguous (and much less like a real English sentence) is

- (4') For all x , lifelong happiness is better than x .

The second sentence cannot be rewritten in these terms because the word “nothing” is not playing the role of a quantifier. (Its nearest mathematical equivalent is something like the *empty set*, that is, the set with no elements.)

Armed with “for all” and “there exists,” we can be clear about the difference between the beginnings of the following sentences.

- (7) Everybody likes at least one drink, namely water.
- (8) Everybody likes at least one drink; I myself go for red wine.

The first sentence makes the point (not necessarily correctly) that there is one drink that everybody likes, whereas the second claims merely that we all have something we like to drink, even if that something varies from person to person. The precise formulations that capture the difference are as follows.

- (7') There exists a drink D such that, for every person P , P likes D .
- (8') For every person P there exists a drink D such that P likes D .

This illustrates an important general principle: if you take a sentence that begins “for every x there exists y such that ...” and interchange the two parts so that it now begins “there exists y such that, for every x , ...,” then you obtain a much stronger statement, since y is no longer allowed to depend on x . If the second statement is still true—that is, if you really can choose a y that works for all the x at once—then the first statement is said to hold *uniformly*.

The symbols \forall and \exists are often used to stand for “for all” and “there exists,” respectively. This allows us to write quite complicated mathematical sentences in a highly symbolic form if we want to. For example, suppose we let P be the set of all primes, as we did earlier. Then the following symbols make the claim that there are infinitely many primes, or rather a slightly different claim that is equivalent to it.

$$(9) \forall n \exists m \quad (m > n) \wedge (m \in P).$$

In words, this says that for every n we can find some m that is both bigger than n and a prime. If we wish to unpack sentence (6) further, we could replace the part $m \in P$ by

$$(10) \forall a, b \quad ab = m \Rightarrow ((a = 1) \vee (b = 1)).$$

There is one final important remark to make about the quantifiers “ \forall ” and “ \exists .” I have presented them as if they

were freestanding, but actually a quantifier is always associated with a set (one says that it *quantifies over* that set). For example, sentence (10) would not be a translation of the sentence “ m is prime” if a and b were allowed to be fractions: if $a = 3$ and $b = \frac{7}{3}$ then $ab = 7$ without either a or b equaling 1, but this does not show that 7 is not a prime. Implicit in the opening symbols $\forall a, b$ is the idea that a and b are intended to be *positive integers*. If this had not been clear from the context, then we could have used the symbol \mathbb{N} (which stands for the set of all positive integers) and started sentence (10) with $\forall a, b \in \mathbb{N}$ instead.

3.3 Negation

The basic idea of negation in mathematics is very simple: there is a symbol, “ \neg ,” which means “not,” and if P is any mathematical statement, then $\neg P$ stands for the statement that is true if and only if P is not true. However, this is another example of a word that has a slightly more restricted meaning to mathematicians than it has in ordinary speech.

To illustrate this phenomenon once again, let us take A to be a set of positive integers and ask ourselves what the negation is of the sentence “Every number in the set A is odd.” Many people when asked this question will suggest, “Every number in the set A is even.” However, this is wrong: if one thinks carefully about what exactly would have to happen for the first sentence to be false, one realizes that all that is needed is that *at least one* number in A should be even. So in fact the negation of the sentence is, “There exists a number in A that is even.”

What explains the temptation to give the first, incorrect answer? One possibility emerges when one writes the sentence more formally, thus:

$$(11) \forall n \in A \quad n \text{ is odd.}$$

The first answer is obtained if one negates just the last part of this sentence, “ n is odd”; but what is asked for is the negation of the *whole sentence*. That is, what is wanted is not

$$(12) \forall n \in A \quad \neg(n \text{ is odd}),$$

but rather

$$(13) \neg(\forall n \in A \quad n \text{ is odd}),$$

which is equivalent to

$$(14) \exists n \in A \quad n \text{ is even.}$$

A second possible explanation is that one is inclined (for psycholinguistic reasons) to think of the phrase “every element of A ” as denoting something like a single, typical element of A . If that comes to have the feel of a particular number n , then we may feel that the negation of “ n is odd” is “ n is even.” The remedy is not to think of the phrase “every element of A ” on its own: it should always be part of the longer phrase, “for every element of A .”

3.4 Free and Bound Variables

Suppose we say something like, “At time t the speed of the projectile is v .” The letters t and v stand for real numbers, and they are called *variables*, because in the back of our mind is the idea that they are changing. More generally, a variable is any letter used to stand for a mathematical object, whether or not one thinks of that object as changing through time. Let us look once again at the formal sentence that said that a positive integer m is prime:

$$(10) \quad \forall a, b \quad ab = m \Rightarrow ((a = 1) \vee (b = 1)).$$

In this sentence, there are three variables, a , b , and m , but there is a very important grammatical and semantic difference between the first two and the third. Here are two results of that difference. First, the sentence does not really make sense unless we already know what m is from the context, whereas it is important that a and b do *not* have any prior meaning. Second, while it makes perfect sense to ask, “For which values of m is sentence (10) true?” it makes no sense at all to ask, “For which values of a is sentence (10) true?” The letter m in sentence (10) stands for a fixed number, not specified in this sentence, while the letters a and b , because of the initial $\forall a, b$, do not *stand for* numbers—rather, in some way they search through all pairs of positive integers, trying to find a pair that multiply together to give m . Another sign of the difference is that you can ask, “What number is m ?” but not, “What number is a ?” A fourth sign is that the meaning of sentence (10) is completely unaffected if one uses different letters for a and b , as in the reformulation

$$(10') \quad \forall c, d \quad cd = m \Rightarrow ((c = 1) \vee (d = 1)).$$

One cannot, however, change m to n without establishing first that n denotes the same integer as m . A variable such as m , which denotes a specific object, is called a *free* variable. It sort of hovers there, free to take any value. A variable like a and b , of the kind that does not denote a specific object, is called a *bound* variable, or sometimes a *dummy* variable. (The word “bound”

is used mainly when the variable appears just after a quantifier, as in sentence (10).)

Yet another indication that a variable is a dummy variable is when the sentence in which it occurs can be rewritten without it. For example, the notation $\sum_{n=1}^{100} f(n)$ is shorthand for $f(1) + f(2) + \cdots + f(100)$, and the second way of writing it does not involve the letter n , so n was not really standing for anything in the first way. Sometimes, actual elimination is not possible, but one feels it could be done in principle. For instance, the sentence “For every real number x , x is either positive, negative, or zero” is a bit like putting together infinitely many sentences such as “ t is either positive, negative, or zero,” one for each real number t , none of which involve a variable.

4 Levels of Formality

It is a surprising fact that a small number of set-theoretic concepts and logical terms can be used to provide a precise language that is versatile enough to express all the statements of ordinary mathematics. There are some technicalities to sort out, but even these can often be avoided if one allows not just sets but also numbers as basic objects. However, if you look at a well-written mathematics paper, then much of it will be written not in symbolic language peppered with symbols such as \forall and \exists , but in what appears to be ordinary English. (Some papers are written in other languages, particularly French, but English has established itself as the international language of mathematics.) How can mathematicians be confident that this ordinary English does not lead to confusion, ambiguity, and even incorrectness?

The answer is that the language typically used is a careful compromise between fully colloquial English, which would indeed run the risk of being unacceptably imprecise, and fully formal symbolism, which would be a nightmare to read. The ideal is to write in as friendly and approachable a way as possible, while making sure that the reader (who, one assumes, has plenty of experience and training in how to read mathematics) can see easily how what one writes could be made more formal if it became important to do so. And sometimes it does become important: when an argument is difficult to grasp it may be that the only way to convince oneself that it is correct is to rewrite it more formally.

Consider, for example, the following reformulation of the principle of mathematical induction, which underlies many proofs:

- (15) Every nonempty set of positive integers has a least element.

If we wish to translate this into a more formal language we need to strip it of words and phrases such as “nonempty” and “has.” But this is easily done. To say that a set A of positive integers is nonempty is simply to say that there is a positive integer that belongs to A . This can be stated symbolically:

$$(16) \exists n \in \mathbb{N} \quad n \in A.$$

What does it mean to say that A has a least element? It means that there exists an element x of A such that every element y of A is either greater than x or equal to x itself. This formulation is again ready to be translated into symbols:

$$(17) \exists x \in A \quad \forall y \in A \quad (y > x) \vee (y = x).$$

Statement (15) says that (16) implies (17) for every set A of positive integers. Thus, it can be written symbolically as follows:

$$(18) \forall A \subset \mathbb{N} \\ [(\exists n \in \mathbb{N} \quad n \in A) \\ \Rightarrow (\exists x \in A \quad \forall y \in A \quad (y > x) \vee (y = x))].$$

Here we have two very different modes of presentation of the same mathematical fact. Obviously (15) is much easier to understand than (18). But if, for example, one is concerned with the foundations of mathematics, or wishes to write a computer program that checks the correctness of proofs, then it is better to work with a greatly pared-down grammar and vocabulary, and then (18) has the advantage. In practice, there are many different levels of formality, and mathematicians are adept at switching between them. It is this that makes it possible to feel completely confident in the correctness of a mathematical argument even when it is *not* presented in the manner of (18)—though it is also this that allows mistakes to slip through the net from time to time.

I.3 Some Fundamental Mathematical Definitions

The concepts discussed in this article occur throughout so much of modern mathematics that it would be inappropriate to discuss them in part III—they are too basic. Many later articles will assume at least some acquaintance with these concepts, so if you have not met them, then reading this article will help you to understand significantly more of the book.

1 The Main Number Systems

Almost always, the first mathematical concept that a child is exposed to is the idea of numbers, and numbers retain a central place in mathematics at all levels. However, it is not as easy as one might think to say what the word “number” means: the more mathematics one learns, the more uses of this word one comes to know, and the more sophisticated one’s concept of number becomes. This individual development parallels a historical development that took many centuries (see FROM NUMBERS TO NUMBER SYSTEMS [II.1]).

The modern view of numbers is that they are best regarded not individually but as parts of larger wholes, called *number systems*; the distinguishing features of number systems are the arithmetical operations—such as addition, multiplication, subtraction, division, and extraction of roots—that can be performed on them. This view of numbers is very fruitful and provides a springboard into abstract algebra. The rest of this section gives a brief description of the five main number systems.

1.1 The Natural Numbers

The *natural numbers*, otherwise known as the *positive integers*, are the numbers familiar even to young children: 1, 2, 3, 4, and so on. It is the natural numbers that we use for the very basic mathematical purpose of counting. The set of all natural numbers is usually denoted \mathbb{N} .

Of course, the phrase “1, 2, 3, 4, and so on” does not constitute a formal definition, but it does suggest the following basic picture of the natural numbers, one that we tend to take for granted.

- (i) Given any natural number n there is another, $n + 1$, that comes next—known as the *successor* of n .
- (ii) A list that starts with 1 and follows each number by its successor will include every natural number exactly once and nothing else.

This picture is encapsulated by THE PEANO AXIOMS [III.69].

Given two natural numbers m and n one can add them together or multiply them, obtaining in each case a new natural number. By contrast, subtraction and division are not always possible. If we want to give meaning to expressions such as $8 - 13$ or $\frac{5}{7}$, then we must work in a larger number system.

1.2 The Integers

The natural numbers are not the only whole numbers, since they do not include zero or negative numbers, both of which are indispensable to mathematics. One of the first reasons for introducing zero was that it is needed for the normal decimal notation of positive integers—how else could one conveniently write 1005? However, it is now thought of as much more than just a convenience, and the property that makes it significant is that it is an *additive identity*, which means that adding zero to any number leaves that number unchanged. And while it is not particularly interesting to do to a number something that has no effect, the property itself is interesting and distinguishes zero from all other numbers. An immediate illustration of this is that it allows us to think about negative numbers: if n is a positive integer, then the defining property of $-n$ is that when you add it to n you get zero.

Somebody with little mathematical experience may unthinkingly assume that numbers are for counting and find negative numbers objectionable because the answer to a question beginning “How many” is never negative. However, simple counting is not the only use for numbers, and there are many situations that are naturally modeled by a number system that includes both positive and negative numbers. For example, negative numbers are sometimes used for the amount of money in a bank account, for temperature (in degrees Celsius or Fahrenheit), and for altitude compared with sea level.

The set of all integers—positive, negative, and zero—is usually denoted \mathbb{Z} (for the German word “Zahlen,” meaning “numbers”). Within this system, subtraction is always possible: that is, if m and n are integers, then so is $m - n$.

1.3 The Rational Numbers

So far we have considered only whole numbers. If we form all possible fractions as well, then we obtain the *rational numbers*. The set of all rational numbers is denoted \mathbb{Q} (for “quotients”).

One of the main uses of numbers besides counting is *measurement*, and most quantities that we measure are ones that can vary continuously, such as length, weight, temperature, and velocity. For these, whole numbers are inadequate.

A more theoretical justification for the rational numbers is that they form a number system in which division is always possible—except by zero. This fact, together with some basic properties of the arithmetical operations, means that \mathbb{Q} is a *field*. What fields are and why

they are important will be explained in more detail later (section 2.2).

1.4 The Real Numbers

A famous discovery of the ancient Greeks, often attributed, despite very inadequate evidence, to the school of PYTHAGORAS [VI.1], was that the square root of 2 is not a rational number. That is, there is no fraction p/q such that $(p/q)^2 = 2$. The Pythagorean theorem about right-angled triangles (which was probably known at least a thousand years before Pythagoras) tells us that if a square has sides of length 1, then the length of its diagonal is $\sqrt{2}$. Consequently, there are lengths that cannot be measured by rational numbers.

This argument seems to give strong practical reasons for extending our number system still further. However, such a conclusion can be resisted: after all, we cannot make any measurements with infinite precision, so in practice we round off to a certain number of decimal places, and as soon as we have done so we have presented our measurement as a rational number. (This point is discussed more fully in NUMERICAL ANALYSIS [IV.20].)

Nevertheless, the *theoretical* arguments for going beyond the rational numbers are irresistible. If we want to solve polynomial equations, take LOGARITHMS [III.25 §4], do trigonometry, or work with the GAUSSIAN DISTRIBUTION [III.73 §5], to give just four examples from an almost endless list, then irrational numbers will appear everywhere we look. They are not used directly for the purposes of measurement, but they are needed if we want to reason theoretically about the physical world by describing it mathematically. This necessarily involves a certain amount of idealization: it is far more convenient to say that the length of the diagonal of a unit square is $\sqrt{2}$ than it is to talk about what would be observed, and with what degree of certainty, if one tried to measure this length as accurately as possible.

The real numbers can be thought of as the set of all numbers with a finite or infinite decimal expansion. In the latter case, they are defined not directly but by a process of successive approximation. For example, the squares of the numbers 1, 1.4, 1.41, 1.414, 1.4142, 1.41421, ..., get as close as you like to 2, if you go far enough along the sequence, which is what we mean by saying that the square root of 2 is the infinite decimal 1.41421....

The set of all real numbers is denoted \mathbb{R} . A more abstract view of \mathbb{R} is that it is an extension of the rational number system to a larger field, and in fact the only one

possible in which processes of the above kind always give rise to numbers that themselves belong to \mathbb{R} .

Because real numbers are intimately connected with the idea of limits (of successive approximations), a true appreciation of the real number system depends on an understanding of mathematical analysis, which will be discussed in section 5.

1.5 The Complex Numbers

Many polynomial equations, such as the equation $x^2 = 2$, do not have rational solutions but can be solved in \mathbb{R} . However, there are many other equations that cannot be solved even in \mathbb{R} . The simplest example is the equation $x^2 = -1$, which has no real solution since the square of any real number is positive or zero. In order to get around this problem, mathematicians introduce a symbol, i , which they treat as a number, and they simply *stipulate* that i^2 is to be regarded as equal to -1 . The *complex number system*, denoted \mathbb{C} , is the set of all numbers of the form $a + bi$, where a and b are real numbers. To add or multiply complex numbers, one treats i as a variable (like x , say), but any occurrences of i^2 are replaced by -1 . Thus,

$$(a + bi) + (c + di) = (a + c) + (b + d)i$$

and

$$\begin{aligned}(a + bi)(c + di) &= ac + bci + adi + bdi^2 \\ &= (ac - bd) + (bc + ad)i.\end{aligned}$$

There are several remarkable points to note about this definition. First, despite its apparently artificial nature, it does not lead to any inconsistency. Secondly, although complex numbers do not directly count or measure anything, they are immensely useful. Thirdly, and perhaps most surprisingly, even though the number i was introduced to help us solve just one equation, it in fact allows us to solve *all* polynomial equations. This is the famous FUNDAMENTAL THEOREM OF ALGEBRA [V.15].

One explanation for the utility of complex numbers is that they provide a concise way to talk about many aspects of geometry, via *Argand diagrams*. These represent complex numbers as points in the plane, the number $a + bi$ corresponding to the point with coordinates (a, b) . If $r = \sqrt{a^2 + b^2}$ and $\theta = \tan^{-1}(b/a)$, then $a = r \cos \theta$ and $b = r \sin \theta$. It turns out that multiplying a complex number $z = x + yi$ by $a + bi$ corresponds to the following geometrical process. First, you associate z with the point (x, y) in the plane. Next, you multiply this point by r , obtaining the point (rx, ry) . Finally, you rotate this new point counterclockwise about the origin through an angle of θ . In other words, the effect

on the complex plane of multiplication by $a + bi$ is to dilate it by r and then rotate it by θ . In particular, if $a^2 + b^2 = 1$, then multiplying by $a + bi$ corresponds to rotating by θ .

For this reason, polar coordinates are at least as good as Cartesian coordinates for representing complex numbers: an alternative way to write $a + bi$ is $re^{i\theta}$, which tells us that the number has distance r from the origin and is positioned at an angle θ around from the positive part of the real axis (in a counterclockwise direction). If $z = re^{i\theta}$ with $r > 0$, then r is called the *modulus* of z , denoted by $|z|$, and θ is the *argument* of z . (Since adding 2π to θ does not change $e^{i\theta}$, it is usually understood that $0 \leq \theta < 2\pi$, or sometimes that $-\pi \leq \theta < \pi$.) One final useful definition: if $z = x + yi$ is a complex number, then its *complex conjugate*, written \bar{z} , is the number $x - yi$. It is easy to check that $z\bar{z} = x^2 + y^2 = |z|^2$.

PUP: Tim wanted to keep this here rather than move it before 'is to dilate' as proofreader suggested.

2 Four Important Algebraic Structures

In the previous section it was emphasized that numbers are best thought of not as individual objects but as members of *number systems*. A number system consists of some objects (numbers) together with operations (such as addition and multiplication) that can be performed on those objects. As such, it is an example of an *algebraic structure*. However, there are many very important algebraic structures that are not number systems, and a few of them will be introduced here.

2.1 Groups

If S is a geometrical shape, then a *rigid motion* of S is a way of moving S in such a way that the distances between the points of S are not changed—squeezing and stretching are not allowed. A rigid motion is a *symmetry* of S if, after it is completed, S looks the same as it did before it moved. For example, if S is an equilateral triangle, then rotating S through 120° about its center is a symmetry; so is reflecting S about a line that passes through one of the vertices of S and the midpoint of the opposite side.

More formally, a symmetry of S is a function f from S to itself such that the distance between any two points x and y of S is the same as the distance between the transformed points $f(x)$ and $f(y)$.

This idea can be hugely generalized: if S is any mathematical structure, then a symmetry of S is a function from S to itself that preserves its structure. If S is a geometrical shape, then the mathematical structure that should be preserved is the distance between any two of

its points. But there are many other mathematical structures that a function may be asked to preserve, most notably algebraic structures of the kind that will soon be discussed. It is fruitful to draw an analogy with the geometrical situation and regard any structure-preserving function as a sort of symmetry.

Because of its extreme generality, symmetry is an all-pervasive concept within mathematics; and wherever symmetries appear, structures known as *groups* follow close behind. To explain what these are and why they appear, let us return to the example of an equilateral triangle, which has, as it turns out, six possible symmetries.

Why is this? Well, let f be a symmetry of an equilateral triangle with vertices A , B , and C and suppose for convenience that this triangle has sides of length 1. Then $f(A)$, $f(B)$, and $f(C)$ must be three points of the triangle and the distances between these points must all be 1. It follows that $f(A)$, $f(B)$, and $f(C)$ are distinct vertices of the triangle, since the furthest apart *any* two points can be is 1 and this happens only when the two points are distinct vertices. So $f(A)$, $f(B)$, and $f(C)$ are the vertices A , B , and C in some order. But the number of possible orders of A , B , and C is 6. It is not hard to show that, once we have chosen $f(A)$, $f(B)$, and $f(C)$, the rest of what f does is completely determined. (For example, if X is the midpoint of A and C , then $f(X)$ must be the midpoint of $f(A)$ and $f(C)$ since there is no other point at distance $\frac{1}{2}$ from $f(A)$ and $f(C)$.)

Let us refer to these symmetries by writing down in order what happens to the vertices A , B , and C . So, for instance, the symmetry ACB is the one that leaves the vertex A fixed and exchanges B and C , which is achieved by reflecting the triangle in the line that joins A to the midpoint of B and C . There are three reflections like this: ACB , CBA , and BAC . There are also two rotations: BCA and CAB . Finally, there is the “trivial” symmetry, ABC , which leaves all points where they were originally. (The “trivial” symmetry is useful in much the same way as zero is useful for the algebra of integer addition.)

What makes these and other sets of symmetries into groups is that any two symmetries can be *composed*, meaning that one symmetry followed by another produces a third (since if two operations both preserve a structure then their combination clearly does too). For example, if we follow the reflection BAC by the reflection ACB , then we obtain the rotation CAB . To work this out, one can either draw a picture or use the following kind of reasoning: the first symmetry takes A to B and the second takes B to C , so the combination takes A to C , and similarly B goes to A , and C to B . Notice that the order

in which we perform the symmetries matters: if we had started with the reflection ACB and then done the reflection BAC , then we would have obtained the rotation BCA . (If you try to see this by drawing a picture, it is important to think of A , B , and C as labels that stay where they are rather than moving with the triangle—they mark positions that the vertices can occupy.)

We can think of symmetries as “objects” in their own right, and of composition as an algebraic operation, a bit like addition or multiplication for numbers. The operation has the following useful properties: it is associative, the trivial symmetry is an identity element, and every symmetry has an inverse. (See BINARY OPERATIONS [I.2 §2.4]. For example, the inverse of a reflection is itself, since doing the same reflection twice leaves the triangle where it started.) More generally, any set with a binary operation that has these properties is called a group. It is *not* part of the definition of a group that the binary operation should be commutative, since, as we have just seen, if one is composing two symmetries then it often makes a difference which one goes first. However, if it is commutative then the group is called *Abelian*, after the Norwegian mathematician Niels Henrik ABEL [VI.32]. The number systems \mathbb{Z} , \mathbb{Q} , \mathbb{R} , and \mathbb{C} all form Abelian groups with the operation of addition, or *under* addition, as one usually says. If you remove zero from \mathbb{Q} , \mathbb{R} , and \mathbb{C} , then they form Abelian groups under multiplication, but \mathbb{Z} does not because of a lack of inverses: the reciprocal of an integer is not usually an integer. Further examples of groups will be given later in this section.

2.2 Fields

Although several number systems form groups, to regard them merely as groups is to ignore a great deal of their algebraic structure. In particular, whereas a group has just one binary operation, the standard number systems have two, namely addition and multiplication (from which further ones, such as subtraction and division, can be derived). The formal definition of a *field* is quite long: it is a set with two binary operations and there are several axioms that these operations must satisfy. Fortunately, there is an easy way to remember these axioms. You just write down all the basic properties you can think of that are satisfied by addition and multiplication in the number systems \mathbb{Q} , \mathbb{R} , and \mathbb{C} .

These properties are as follows. Both addition and multiplication are commutative and associative, and both have identity elements (0 for addition and 1 for multiplication). Every element x has an additive inverse $-x$ and a multiplicative inverse $1/x$ (except that 0 does

not have a multiplicative inverse). It is the existence of these inverses that allows us to define subtraction and division: $x - y$ means $x + (-y)$ and x / y means $x \cdot (1/y)$.

That covers all the properties that addition and multiplication satisfy individually. However, a very general rule when defining mathematical structures is that if a definition splits into parts, then the definition as a whole will not be interesting *unless those parts interact*. Here our two parts are addition and multiplication, and the properties mentioned so far do not relate them in any way. But one final property, known as the *distributive law*, does this, and thereby gives fields their special character. This is the rule that tells us how to multiply out brackets: $x(y + z) = xy + xz$ for any three numbers x , y , and z .

Having listed these properties, one may then view the whole situation *abstractly* by regarding the properties as axioms and saying that a field is any set with two binary operations that satisfy all those axioms. However, when one works in a field, one usually thinks of the axioms not as a list of statements but rather as a general license to do all the algebraic manipulations that one can do when talking about rational, real, and complex numbers.

Clearly, the more axioms one has, the harder it is to find a mathematical structure that satisfies them, and it is indeed the case that fields are harder to come by than groups. For this reason, the best way to understand fields is probably to concentrate on examples. In addition to \mathbb{Q} , \mathbb{R} , and \mathbb{C} , one other field stands out as fundamental, namely \mathbb{F}_p , which is the set of integers modulo a prime p , with addition and multiplication also defined modulo p (see MODULAR ARITHMETIC [III.60]).

What makes fields interesting, however, is not so much the existence of these basic examples as the fact that there is an important process of *extension* that allows one to build new fields out of old ones. The idea is to start with a field \mathbb{F} , find a polynomial P that has no roots in \mathbb{F} , and “adjoin” a new element to \mathbb{F} with the stipulation that it is a root of P . This produces an extended field \mathbb{F}' , which consists of everything that one can produce from this root and from elements of \mathbb{F} using addition and multiplication.

We have already seen an important example of this process: in the field \mathbb{R} , the polynomial $P(x) = x^2 + 1$ has no root, so we adjoined the element i and let \mathbb{C} be the field of all combinations of the form $a + bi$.

We can apply exactly the same process to the field \mathbb{F}_3 , in which again the equation $x^2 + 1 = 0$ has no solution. If we do so, then we obtain a new field, which, like \mathbb{C} , consists of all combinations of the form $a + bi$, but now a and b belong to \mathbb{F}_3 . Since \mathbb{F}_3 has three elements,

this new field has nine elements. Another example is the field $\mathbb{Q}(\sqrt{2})$, which consists of all numbers of the form $a + b\sqrt{2}$, where now a and b are rational numbers. A slightly more complicated example is $\mathbb{Q}(y)$, where y is a root of the polynomial $x^3 - x - 1$. A typical element of this field has the form $a + by + cy^2$, with a , b , and c rational. If one is doing arithmetic in $\mathbb{Q}(y)$, then whenever y^3 appears, it can be replaced by $y + 1$ (because $y^3 - y - 1 = 0$), just as i^2 can be replaced by -1 in the complex numbers. For more on why field extensions are interesting, see the discussion of AUTOMORPHISMS in section 4.1.

A second very significant justification for introducing fields is that they can be used to form vector spaces, and it is to these that we now turn.

2.3 Vector Spaces

One of the most convenient ways to represent points in a plane that stretches out to infinity in all directions is to use Cartesian coordinates. One chooses an origin and two directions X and Y , usually at right angles to each other. Then the pair of numbers (a, b) stands for the point you reach in the plane if you go a distance a in direction X and a distance b in direction Y (where if a is a negative number such as -2 , this is interpreted as going a distance $+2$ in the opposite direction to X , and similarly for b).

Another way of saying the same thing is this. Let \mathbf{x} and \mathbf{y} stand for the unit vectors in directions X and Y , respectively, so their Cartesian coordinates are $(1, 0)$ and $(0, 1)$. Then every point in the plane is a so-called *linear combination* $a\mathbf{x} + b\mathbf{y}$ of the *basis vectors* \mathbf{x} and \mathbf{y} . To interpret the expression $a\mathbf{x} + b\mathbf{y}$, first rewrite it as $a(1, 0) + b(0, 1)$. Then a times the unit vector $(1, 0)$ is $(a, 0)$ and b times the unit vector $(0, 1)$ is $(0, b)$ and when you add $(a, 0)$ and $(0, b)$ coordinate by coordinate you get the vector (a, b) .

Here is another situation where linear combinations appear. Suppose you are presented with the differential equation $(d^2y/dx^2) + y = 0$, and happen to know (or notice) that $y = \sin x$ and $y = \cos x$ are two possible solutions. Then you can easily check that $y = a \sin x + b \cos x$ is a solution for any pair of numbers a and b . That is, any linear combination of the existing solutions $\sin x$ and $\cos x$ is another solution. It turns out that all solutions are of this form, so we can regard $\sin x$ and $\cos x$ as “basis vectors” for the “space” of solutions of the differential equation.

Linear combinations occur in many many contexts throughout mathematics. To give one more example,

PUP: Tim would like to keep 'brackets' as even he, as a mathematician, would say 'brackets' rather than the more formal 'parentheses'. OK?

PUP: Tim and I both think this cross-referencing sentence works well but I wanted to draw your attention to it in case you weren't so happy with it. There aren't many cross-references like this in the volume.

an arbitrary polynomial of degree 3 has the form $ax^3 + bx^2 + cx + d$, which is a linear combination of the four basic polynomials 1, x , x^2 , and x^3 .

A *vector space* is a mathematical structure in which the notion of linear combination makes sense. The objects that belong to the vector space are usually called *vectors*, unless we are talking about a specific example and are thinking of them as concrete objects such as polynomials or solutions of a differential equation. Slightly more formally, a vector space is a set V such that, given any two vectors \mathbf{v} and \mathbf{w} (that is, elements of V) and any two real numbers a and b , we can form the linear combination $a\mathbf{v} + b\mathbf{w}$.

Notice that this linear combination involves objects of two different kinds, the vectors \mathbf{v} and \mathbf{w} and the numbers a and b . The latter are known as *scalars*. The operation of forming linear combinations can be broken up into two constituent parts: addition and scalar multiplication. To form the combination $a\mathbf{v} + b\mathbf{w}$, first multiply the vectors \mathbf{v} and \mathbf{w} by the scalars a and b , obtaining the vectors $a\mathbf{v}$ and $b\mathbf{w}$, and then add these resulting vectors to obtain the full combination $a\mathbf{v} + b\mathbf{w}$.

The definition of linear combination must obey certain natural rules. Addition of vectors must be commutative and associative, with an identity, the *zero vector*, and inverses for each \mathbf{v} (written $-\mathbf{v}$). Scalar multiplication must obey a sort of associative law, namely that $a(b\mathbf{v})$ and $(ab)\mathbf{v}$ are always equal. We also need two distributive laws: $(a + b)\mathbf{v} = a\mathbf{v} + b\mathbf{v}$ and $a(\mathbf{v} + \mathbf{w}) = a\mathbf{v} + a\mathbf{w}$ for any scalars a and b and any vectors \mathbf{v} and \mathbf{w} .

Another context in which linear combinations arise, one that lies at the heart of the usefulness of vector spaces, is the solution of simultaneous equations. Suppose one is presented with the two equations $3x + 2y = 6$ and $x - y = 7$. The usual way to solve such a pair of equations is to try to eliminate either x or y by adding an appropriate multiple of one of the equations to the other: that is, by taking a certain linear combination of the equations. In this case, we can eliminate y by adding twice the second equation to the first, obtaining the equation $5x = 20$, which tells us that $x = 4$ and hence that $y = -3$. Why were we allowed to combine equations like this? Well, let us write L_1 and R_1 for the left- and right-hand sides of the first equation, and similarly L_2 and R_2 for the second. If, for some particular choice of x and y , it is true that $L_1 = R_1$ and $L_2 = R_2$, then clearly $L_1 + 2L_2 = R_1 + 2R_2$, as the two sides of this equation are merely giving different names to the same numbers.

Given a vector space V , a *basis* is a collection of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ with the following property: every vector

in V can be written in exactly one way as a linear combination $a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \dots + a_n\mathbf{v}_n$. There are two ways in which this can fail: there may be a vector that cannot be written as a linear combination of $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ or there may be a vector that can be so expressed, but in more than one way. If every vector is a linear combination then we say that the vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ *span* V , and if no vector is a linear combination in more than one way then we say that they are *independent*. An equivalent definition is that $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ are independent if the only way of writing the zero vector as $a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \dots + a_n\mathbf{v}_n$ is by taking $a_1 = a_2 = \dots = a_n = 0$.

The number of elements in a basis is called the *dimension* of V . It is not immediately obvious that there could not be two bases of different sizes, but it turns out that there cannot, so the concept of dimension makes sense. For the plane, the vectors \mathbf{x} and \mathbf{y} defined earlier formed a basis, so the plane, as one would hope, has dimension 2. If we were to take more than two vectors, then they would no longer be independent: for example, if we take the vectors $(1, 2)$, $(1, 3)$, and $(3, 1)$, then we can write $(0, 0)$ as the linear combination $8(1, 2) - 5(1, 3) - (3, 1)$. (To work this out one must solve some simultaneous equations—this is typical of calculations in vector spaces.)

The most obvious n -dimensional vector space is the space of all sequences (x_1, \dots, x_n) of n real numbers. To add this to a sequence (y_1, \dots, y_n) one simply forms the sequence $(x_1 + y_1, \dots, x_n + y_n)$ and to multiply it by a scalar c one forms the sequence (cx_1, \dots, cx_n) . This vector space is denoted \mathbb{R}^n . Thus, the plane with its usual coordinate system is \mathbb{R}^2 and three-dimensional space is \mathbb{R}^3 .

It is not in fact necessary for the number of vectors in a basis to be finite. A vector space that does not have a finite basis is called *infinite dimensional*. This is not an exotic property: many of the most important vector spaces, particularly spaces where the “vectors” are functions, are infinite dimensional.

There is one final remark to make about scalars. They were defined earlier as real numbers that one uses to make linear combinations of vectors. But it turns out that the calculations one does with scalars, in particular solving simultaneous equations, can all be done in a more general context. What matters is that they should belong to a field, so \mathbb{Q} , \mathbb{R} , and \mathbb{C} can all be used as systems of scalars, as indeed can more general fields. If the scalars for a vector space V come from a field \mathbb{F} , then one says that V is a vector space *over* \mathbb{F} . This generalization is important and useful: see, for example, ALGEBRAIC NUMBERS [IV.3 §17].

2.4 Rings

Another algebraic structure that is very important is a *ring*. Rings are not quite as central to mathematics as groups, fields, or vector spaces, so a proper discussion of them will be deferred to RINGS, IDEALS, AND MODULES [III.82]. However, roughly speaking, a ring is an algebraic structure that has most, but not necessarily all, of the properties of a field. In particular, the requirements of the multiplicative operation are less strict. The most important relaxation is that nonzero elements of a ring are not required to have multiplicative inverses; but sometimes multiplication is not even required to be commutative. If it is, then the ring itself is said to be commutative—a typical example of a commutative ring is the set \mathbb{Z} of all integers. Another is the set of all polynomials with coefficients in some field \mathbb{F} .

3 Creating New Structures Out of Old Ones

An important first step in understanding the definition of some mathematical structure is to have a supply of examples. Without examples, a definition is dry and abstract. With them, one begins to have a feeling for the structure that its definition alone cannot usually provide.

One reason for this is that it makes it much easier to answer basic questions. If you have a general statement about structures of a given type and want to know whether it is true, then it is very helpful if you can test it in a wide range of particular cases. If it passes all the tests, then you have some evidence in favor of the statement. If you are lucky, you may even be able to see why it is true; alternatively, you may find that the statement is true for each example you try, but always for reasons that depend on particular features of the example you are examining. Then you will know that you should try to avoid these features if you want to find a counterexample. If you *do* find a counterexample, then the general statement is false, but it may still happen that a modification to the statement is true and useful. In that case, the counterexample will help you to find an appropriate modification.

The moral, then, is that examples are important. So how does one find them? There are two completely different approaches. One is to build them from scratch. For example, one might define a group G to be the group of all symmetries of an icosahedron. Another, which is the main topic of this section, is to take some already constructed examples and build new ones out of them. For example, the group \mathbb{Z}^2 , which consists of all pairs of integers (x, y) , with addition defined by the obvious

rule $(x, y) + (x', y') = (x + x', y + y')$, is a “product” of two copies of the group \mathbb{Z} . As we shall see, this notion of product is very general and can be applied in many other contexts. But first let us look at an even more basic method of finding new examples.

3.1 Substructures

As we saw earlier, the set \mathbb{C} of all complex numbers, with the operations of addition and multiplication, forms one of the most basic examples of a field. It also contains many *subfields*: that is, subsets that themselves form fields. Take, for example, the set $\mathbb{Q}(i)$ of all complex numbers of the form $a + bi$ for which a and b are rational. This is a subset of \mathbb{C} and is also a field. To show this, one must prove that $\mathbb{Q}(i)$ is *closed* under addition, multiplication, and the taking of inverses. That is, if z and w are elements of $\mathbb{Q}(i)$, then $z + w$ and zw must be as well, as must $-z$ and $1/z$ (this last requirement applying only when $z \neq 0$). Axioms such as the commutativity and associativity of addition and multiplication are then true in $\mathbb{Q}(i)$ for the simple reason that they are true in the larger set \mathbb{C} .

Even though $\mathbb{Q}(i)$ is contained in \mathbb{C} , it is a more interesting field in some important ways. But how can this be? Surely, one might think, an object cannot become *more* interesting when most of it is taken away. But a moment's further thought shows that it certainly can: for example, the set of all prime numbers contains fascinating mysteries of a kind that one does not expect to encounter in the set of all positive integers. As for fields, THE FUNDAMENTAL THEOREM OF ALGEBRA [V.15] tells us that every polynomial equation has a solution in \mathbb{C} . This is very definitely not true in $\mathbb{Q}(i)$. So in $\mathbb{Q}(i)$, and in many other fields of a similar kind, we can ask which polynomial equations have solutions. This turns out to be a deep and important question that simply does not arise in the larger field \mathbb{C} .

In general, given an example X of an algebraic structure, a substructure of X is a subset Y that has relevant closure properties. For instance, groups have subgroups, vector spaces have subspaces, rings have subrings (and also IDEALS [III.82]), and so on. If the property defining the substructure Y is a sufficiently interesting one, then Y may well be significantly different from X and may therefore be a useful addition to one's stock of examples.

This discussion has focused on algebra, but interesting substructures abound in analysis and geometry as well. For example, the plane \mathbb{R}^2 is not a particularly interesting set, but it has subsets, such as the MANDELBROT

SET [IV.15 §2.8], to give just one example, that are still far from fully understood.

3.2 Products

Let G and H be two groups. The *product group* $G \times H$ has as its elements all pairs of the form (g, h) such that g belongs to G and h belongs to H . This definition shows how to build the elements of $G \times H$ out of the elements of G and the elements of H . But to define a group we need to do more: we are given binary operations on G and H and we must use them to build a binary operation on $G \times H$. If g_1 and g_2 are elements of G , let us write $g_1 g_2$ for the result of applying G 's binary operation to them, as is customary, and let us do the same for H . Then there is an obvious binary operation we can define on the pairs, namely

$$(g_1, h_1)(g_2, h_2) = (g_1 g_2, h_1 h_2).$$

That is, one applies the binary operation from G to the first coordinate and the binary operation from H to the second.

One can form products of vector spaces in a very similar way. If V and W are two vector spaces, then the elements of $V \times W$ are all pairs of the form (v, w) with v in V and w in W . Addition and scalar multiplication are defined by the formulas

$$(v_1, w_1) + (v_2, w_2) = (v_1 + v_2, w_1 + w_2)$$

and

$$\lambda(v, w) = (\lambda v, \lambda w).$$

The dimension of the resulting space is the sum of the dimensions of V and W . (It is actually more usual to denote this space by $V \oplus W$ and call it the *direct sum* of V and W . Nevertheless, it is a product construction.)

It is not always possible to define product structures in this simple way. For example, if \mathbb{F}_1 and \mathbb{F}_2 are two fields, we might be tempted to define a "product field" $\mathbb{F}_1 \times \mathbb{F}_2$ using the formulas

$$(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2)$$

and

$$(x_1, y_1)(x_2, y_2) = (x_1 x_2, y_1 y_2).$$

However, with this definition we do not obtain a field. Most of the axioms hold, including the existence of additive and multiplicative identities—they are $(0, 0)$ and $(1, 1)$, respectively—but the nonzero element $(1, 0)$ does not have a multiplicative inverse, since $(1, 0)(x, y) = (x, 0)$, which can never equal $(1, 1)$.

Occasionally we can define more complicated binary operations that do make the set $\mathbb{F}_1 \times \mathbb{F}_2$ into a field. For

instance, if $\mathbb{F}_1 = \mathbb{F}_2 = \mathbb{R}$, then we can define addition as above, but define multiplication in a less obvious way as follows:

$$(x_1, y_1)(x_2, y_2) = (x_1 x_2 - y_1 y_2, x_1 y_2 + x_2 y_1).$$

Then we obtain \mathbb{C} , the field of complex numbers, since the pair (x, y) can be identified with the complex number $x + iy$. However, this is not a product field in the general sense we are discussing.

Returning to groups, what we defined earlier was the *direct product* of G and H . However, there are other, more complicated products of groups, which can be used to give a much richer supply of examples. To illustrate this, let us consider the *dihedral group* D_4 , which is the group of all symmetries of a square, of which there are eight. If we let R stand for one of the reflections and T for a counterclockwise quarter turn, then every symmetry can be written in the form $T^i R^j$, where i is 0, 1, 2, or 3 and j is 0 or 1. (Geometrically, this says that you can produce any symmetry by either rotating through a multiple of 90° or reflecting and then rotating.)

This suggests that we might be able to regard D_4 as a product of the group $\{I, T, T^2, T^3\}$, consisting of four rotations, with the group $\{I, R\}$, consisting of the identity I and the reflection R . We could even write (T^i, R^j) instead of $T^i R^j$. However, we have to be careful. For instance, $(TR)(TR)$ does not equal $T^2 R^2 = T^2$ but I . The correct rule for multiplication can be deduced from the fact that $RTR = T^{-1}$ (which in geometrical terms is saying that if you reflect the square, rotate it counterclockwise through 90° , and reflect back, then the result is a *clockwise* rotation through 90°). It turns out to be

$$(T^i, R^j)(T^{i'}, R^{j'}) = (T^{i-i'}, R^{j+j'}).$$

For example, the product of (T, R) with (T^3, R) is $T^{-2} R^2$, which equals T^2 .

This is a simple example of a "semi-direct product" of two groups. In general, given two groups G and H , there may be several interesting ways of defining a binary operation on the set of pairs (g, h) , and therefore several potentially interesting new groups.

3.3 Quotients

Let us write $\mathbb{Q}[x]$ for the set of all polynomials in the variable x with rational coefficients: that is, expressions like $2x^4 - \frac{3}{2}x + 6$. Any two such polynomials can be added, subtracted, or multiplied together and the result will be another polynomial. This makes $\mathbb{Q}[x]$ into a commutative ring, but not a field, because if you divide one polynomial by another then the result is not (necessarily) a polynomial.

We will now convert $\mathbb{Q}[x]$ into a field in what may at first seem a rather strange way: by regarding the polynomial $x^3 - x - 1$ as “equivalent” to the zero polynomial. To put this another way, whenever a polynomial involves x^3 we will allow ourselves to replace x^3 by $x + 1$, and we will regard the new polynomial that results as equivalent to the old one. For example, writing “ \sim ” for “is equivalent to”:

$$\begin{aligned} x^5 &= x^3 x^2 \sim (x + 1)x^2 = x^3 + x^2 \\ &\sim x + 1 + x^2 = x^2 + x + 1. \end{aligned}$$

Notice that in this way we can convert any polynomial into one of degree at most 2, since whenever the degree is higher, you can reduce it by taking out x^3 from the term of highest degree and replacing it by $x + 1$, just as we did above.

Notice also that whenever we do such a replacement, the difference between the old polynomial and the new one is a multiple of $x^3 - x - 1$. For example, when we replaced $x^3 x^2$ by $(x + 1)x^2$ the difference was $(x^3 - x - 1)x^2$. Therefore, what our process amounts to is this: two polynomials are equivalent if and only if their difference is a multiple of the polynomial $x^3 - x - 1$.

Now the reason $\mathbb{Q}[x]$ was not a field was that nonconstant polynomials do not have multiplicative inverses. For example, it is obvious that one cannot multiply x^2 by a polynomial and obtain the polynomial 1. However, we can obtain a polynomial that is *equivalent* to 1 if we multiply by $1 + x - x^2$. Indeed, the product of the two is

$$x^2 + x^3 - x^4 \sim x^2 + x + 1 - (x + 1)x = 1.$$

It turns out that *all* polynomials that are not equivalent to zero (that is, are not multiples of $x^3 - x - 1$) have multiplicative inverses in this generalized sense. (To find an inverse for a polynomial P one applies the generalized EUCLID ALGORITHM [III.22] to find polynomials Q and R such that $PQ + R(x^3 - x - 1) = 1$. The reason we obtain 1 on the right-hand side is that $x^3 - x - 1$ cannot be factorized in $\mathbb{Q}[x]$ and P is not a multiple of $x^3 - x - 1$, so their highest common factor is 1. The inverse of P is then Q .)

In what sense does this mean that we have a field? After all, the product of x^2 and $1 + x - x^2$ was not 1: it was merely equivalent to 1. This is where the notion of quotients comes in. We simply decide that when two polynomials are equivalent, we will regard them as equal, and we denote the resulting mathematical structure by $\mathbb{Q}[x]/(x^3 - x - 1)$. This structure turns out to be a field, and it turns out to be important as the smallest field that contains \mathbb{Q} and also has a root of the polynomial $X^3 - X - 1$. What is this root? It is simply x .

This is a slightly subtle point because we are now thinking of polynomials in two different ways: as elements of $\mathbb{Q}[x]/(x^3 - x - 1)$ (at least when equivalent ones are regarded as equal), and also as functions defined on $\mathbb{Q}[x]/(x^3 - x - 1)$. So the polynomial $X^3 - X - 1$ is not the zero polynomial, since for example it takes the value 5 when $X = 2$ and the value $x^6 - x^2 - 1 \sim (x + 1)^2 - x^2 - 1 \sim 2x$ when $X = x^2$.

You may have noticed a strong similarity between the discussion of the field $\mathbb{Q}[x]/(x^3 - x - 1)$ and the discussion of the field $\mathbb{Q}(y)$ at the end of section 2.2. And indeed, this is no coincidence: they are two different ways of describing the same field. However, thinking of the field as $\mathbb{Q}/(x^3 - x - 1)$ brings significant advantages, as it converts questions about a mysterious set of complex numbers into more approachable questions about polynomials.

What does it mean to “regard two mathematical objects as equal” when they are not equal? A formal answer to this question uses the notion of equivalence relations and equivalence classes (discussed in THE LANGUAGE AND GRAMMAR OF MATHEMATICS [I.2 §2.3]): one says that the elements of $\mathbb{Q}[x]/(x^3 - x - 1)$ are not in fact polynomials but *equivalence classes* of polynomials. However, to understand the notion of a quotient it is much easier to look at an example with which we are all familiar, namely the set \mathbb{Q} of rational numbers. If we are trying to explain carefully what a rational number is, then we may start by saying that a typical rational number has the form a/b , where a and b are integers and b is not 0. And it is possible to define the set of rational numbers to be the set of all such expressions, with the rules

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}$$

and

$$\frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}.$$

However, there is one very important further remark we must make, which is that we do not regard all such expressions as different: for example, $\frac{1}{2}$ and $\frac{3}{6}$ are supposed to be the same rational number. So we define two expressions $\frac{a}{b}$ and $\frac{c}{d}$ to be equivalent if $ad = bc$ and we regard equivalent expressions as denoting the same number. Notice that the expressions can be genuinely different, but we think of them as denoting the same object.

If we do this, then we must be careful whenever we define functions and binary operations. For example, suppose we tried to define a binary operation “ \circ ” on \mathbb{Q}

by the natural-looking formula

$$\frac{a}{b} \circ \frac{c}{d} = \frac{a+c}{b+d}.$$

This definition turns out to have a very serious flaw. To see why, let us apply it to the fractions $\frac{1}{2}$ and $\frac{1}{3}$. Then it gives us the answer $\frac{2}{5}$. Now let us replace $\frac{1}{2}$ by the equivalent fraction $\frac{3}{6}$ and apply the formula again. This time it gives us the answer $\frac{4}{9}$, which is different. Thus, although the formula defines a perfectly good binary operation on the set of *expressions* of the form $\frac{a}{b}$, it does not make any sense as a binary operation on the set of *rational numbers*.

In general, it is essential to check that if you put equivalent objects in then you get equivalent objects out. For example, when defining addition and multiplication for the field $\mathbb{Q}[x]/(x^3 - x - 1)$, one must check that if P and P' differ by a multiple of $x^3 - x - 1$, and Q and Q' also differ by a multiple of $x^3 - x - 1$, then so do $P + Q$ and $P' + Q'$, and so do PQ and $P'Q'$. This is an easy exercise.

Why is the word “quotient” used? Well, a quotient is normally what you get when you divide one number by another, so to understand the analogy let us think about dividing 21 by 3. We can think of this as dividing up twenty-one objects into sets of three objects each and asking how many sets we get. This can be described in terms of equivalence as follows. Let us call two objects equivalent if they belong to the same one of the seven sets. Then there can be at most seven inequivalent objects. So when we regard equivalent objects as the same, we “divide out by the equivalence,” obtaining a “quotient set” that has seven elements.

A rather different use of quotients leads to an elegant definition of the mathematical shape known as a *torus*: that is, the shape of the surface of a doughnut (of the kind that has a hole). We start with the plane, \mathbb{R}^2 , and define two points (x, y) and (x', y') to be equivalent if $x - x'$ and $y - y'$ are both integers. Suppose that we regard any two equivalent points as the same and that we start at a point (x, y) and move right until we reach the point $(x + 1, y)$. This point is “the same” as (x, y) , since the difference is $(1, 0)$. Therefore, it is as though the entire plane has been wrapped around a vertical cylinder of circumference 1 and we have gone around this cylinder once. If we now apply the same argument to the y -coordinate, noting that (x, y) is always “the same” point as $(x, y + 1)$, then we find that this cylinder is itself “folded around” so that if you go “upwards” by a distance of 1 then you get back to where you started. But that is what a torus is: a cylinder that is folded back into itself. (This is not the only way of defining a torus,

however. For example, it can be defined as the product of two circles.)

Many other important objects in modern geometry are defined using quotients. It often happens that the object one starts with is extremely big, but that at the same time the equivalence relation is very generous, in the sense that it is easy for one object to be equivalent to another. In that case the number of “genuinely distinct” objects can be quite small. This is a rather loose way of talking, since it is not really the *number* of distinct objects that is interesting so much as the complexity of the set of these objects. It might be better to say that one often starts with a hopelessly large and complicated structure but “divides out most of the mess” and ends up with a quotient object that has a structure that is simple enough to be manageable while still conveying important information. Good examples of this are the FUNDAMENTAL GROUP [IV.10 §3] and the HOMOLOGY AND COHOMOLOGY GROUPS [IV.10 §2] of a topological space; an even better example is the notion of a MODULI SPACE [IV.8].

Many people find the idea of a quotient somewhat difficult to grasp, but it is of major importance throughout mathematics, which is why it has been discussed at some length here.

4 Functions between Algebraic Structures

One rule with almost no exceptions is that mathematical structures are not studied in isolation: as well as the structures themselves one looks at certain *functions* defined on those structures. In this section we shall see which functions are worth considering, and why. (For a discussion of functions in general, see THE LANGUAGE AND GRAMMAR OF MATHEMATICS [I.2 §2.2].)

4.1 Homomorphisms, Isomorphisms, and Automorphisms

If X and Y are two examples of a particular mathematical structure, such as a group, field, or vector space, then, as was suggested in the discussion of symmetry in section 2.1, there is a class of functions from X to Y of particular interest, namely the functions that “preserve the structure.” Roughly speaking, a function $f: X \rightarrow Y$ is said to preserve the structure of X if, given any relationship between elements of X that is expressed in terms of that structure, there is a corresponding relationship between the images of those elements that is expressed in terms of the structure of Y . For example, if X and Y are groups and a, b , and c are elements of X such that $ab = c$, then, if f is to preserve the algebraic structure of X , $f(a)f(b)$ must equal $f(c)$ in Y . (Here, as is usual,

we are using the same notation for the binary operations that make X and Y groups as is normally used for multiplication.) Similarly, if X and Y are fields, with binary operations that we shall write using the standard notation for addition and multiplication, then a function $f : X \rightarrow Y$ will be interesting only if $f(a) + f(b) = f(c)$ whenever $a + b = c$, and $f(a)f(b) = f(c)$ whenever $ab = c$. For vector spaces, the functions of interest are ones that preserve linear combinations: if V and W are vector spaces, then $f(av + bw)$ should always equal $af(v) + bf(w)$.

A function that preserves structure is generally known as a *homomorphism*, though homomorphisms of particular mathematical structures often have their own names: for example, a homomorphism of vector spaces is called a linear map.

There are some useful properties that a homomorphism may have if we are lucky. To see why further properties can be desirable, consider the following example. Let X and Y be groups and let $f : X \rightarrow Y$ be the function that takes every element of X to the identity element e of Y . Then, according to the definition above, f preserves the structure of X , since whenever $ab = c$, we have $f(a)f(b) = ee = e = f(c)$. However, it seems more accurate to say that f has *collapsed* the structure. One can make this idea more precise: although $f(a)f(b) = f(c)$ whenever $ab = c$, the *converse does not hold*: it is perfectly possible for $f(a)f(b)$ to equal $f(c)$ without ab equaling c , and indeed that happens in the example just given.

An *isomorphism* between two structures X and Y is a homomorphism $f : X \rightarrow Y$ that has an inverse $g : Y \rightarrow X$ that is also a homomorphism. For most algebraic structures, if f has an inverse g , then g is automatically a homomorphism; in such cases we can simply say that an isomorphism is a homomorphism that is also a **BIJECTION** [I.2 §2.2]. That is, f is a one-to-one correspondence between X and Y that preserves structure.¹

If X and Y are fields, then these considerations are less interesting: it is a simple exercise to show that every homomorphism $f : X \rightarrow Y$ is automatically an isomorphism between X and its image $f(X)$, that is, the set of all values taken by the function f . So structure cannot

be collapsed without being lost. (The proof depends on the fact that the zero in Y has no multiplicative inverse.)

In general, if there is an isomorphism between two algebraic structures X and Y , then X and Y are said to be *isomorphic* (coming from the Greek words for “same” and “shape”). Loosely, the word “isomorphic” means “the same in all essential respects,” where what counts as essential is precisely the algebraic structure. What is absolutely *not* essential is the nature of the objects that have the structure: for example, one group might consist of certain complex numbers, another of integers modulo a prime p , and a third of rotations of a geometrical figure, and they could all turn out to be isomorphic. The idea that two mathematical constructions can have very different constituent parts and yet in a deeper sense be “the same” is one of the most important in mathematics.

An *automorphism* of an algebraic structure X is an isomorphism from X to itself. Since it is hardly surprising that X is isomorphic to itself, one might ask what the point is of automorphisms. The answer is that automorphisms are precisely the algebraic symmetries alluded to in our discussion of groups. An automorphism of X is a function from X to itself that preserves the structure (which now comes in the form of statements like $ab = c$). The composition of two automorphisms is clearly a third, and as a result the automorphisms of a structure X form a group. Although the individual automorphisms may not be of much interest, the group certainly is, as it often encapsulates what one really wants to know about a structure X that is too complicated to analyze directly.

A spectacular example of this is when X is a field. To illustrate, let us take the example of $\mathbb{Q}(\sqrt{2})$. If $f : \mathbb{Q}(\sqrt{2}) \rightarrow \mathbb{Q}(\sqrt{2})$ is an automorphism, then $f(1) = 1$, as we have seen, and then $f(2) = f(1+1) = f(1) + f(1) = 1 + 1 = 2$. Continuing like this, we can show that $f(n) = n$ for every positive integer n . Then $f(n) + f(-n) = f(n + (-n)) = f(0) = 0$, so $f(-n) = -f(n) = -n$. Finally, $f(p/q) = f(p)/f(q) = p/q$ when p and q are integers with $q \neq 0$. So f takes every rational number to itself. What can we say about $f(\sqrt{2})$? Well, $f(\sqrt{2})f(\sqrt{2}) = f(\sqrt{2} \cdot \sqrt{2}) = f(2) = 2$, but this implies only that $f(\sqrt{2})$ is $\sqrt{2}$ or $-\sqrt{2}$. It turns out that both choices are possible: one automorphism is the “trivial” one $f(a + b\sqrt{2}) = a + b\sqrt{2}$ and the other is the more interesting one $f(a + b\sqrt{2}) = a - b\sqrt{2}$. This observation demonstrates that there is no algebraic difference between the two square roots; in this sense, the field $\mathbb{Q}(\sqrt{2})$ does not know which square root of 2 is positive and which negative. These two automorphisms form a group, which is isomorphic to the group consisting of

PUP: large footnote created here, as the discussion that completed this paragraph before was more footnote-like in character. OK?

1. Let us see how this claim is proved for groups. If X and Y are groups, $f : X \rightarrow Y$ is a homomorphism with inverse $g : Y \rightarrow X$ and u, v , and w are elements of Y with $uv = w$, then we must show that $g(u)g(v) = g(w)$. To do this, let $a = g(u)$, $b = g(v)$, and $d = g(w)$. Since f and g are inverse functions, $f(a) = u$, $f(b) = v$, and $f(d) = w$. Now let $c = ab$. Then $w = uv = f(a)f(b) = f(c)$, since f is a homomorphism. But then $f(c) = f(d)$, which implies that $c = d$ (just apply the function g to $f(c)$ and $f(d)$). Therefore $ab = d$, which tells us that $g(u)g(v) = g(w)$, as we needed to show.

the elements ± 1 under multiplication, or the group of integers modulo 2, or the group of symmetries of an isosceles triangle that is not equilateral, or The list is endless.

The automorphism groups associated with certain field extensions are called GALOIS GROUPS [III.30], and are a vital component of the proof of THE INSOLUBILITY OF THE QUINTIC [V.24], as well as of large parts of algebraic number theory (see ALGEBRAIC NUMBERS [IV.3]).

4.2 Linear Maps and Matrices

Homomorphisms between vector spaces have a distinctive geometrical property: they send straight lines to straight lines. For this reason they are called *linear maps*, as was mentioned in the previous subsection. From a more algebraic point of view, the structure that linear maps preserve is that of linear combinations: a function f from one vector space to another is a linear map if $f(a\mathbf{u} + b\mathbf{v}) = af(\mathbf{u}) + bf(\mathbf{v})$ for every pair of vectors $\mathbf{u}, \mathbf{v} \in V$ and every pair of scalars a and b . From this one can deduce the more general assertion that $f(a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n)$ is always equal to $a_1f(\mathbf{v}_1) + \cdots + a_nf(\mathbf{v}_n)$.

Suppose that we wish to define a linear map from V to W . How much information do we need to provide? This may seem a vague question, so here is a similar one. How much information is needed to specify a point in space? The answer is that, once one has devised a sensible coordinate system, three numbers will suffice. If the point is not too far from Earth's surface then one might wish to use its latitude, its longitude, and its height above sea level, for instance. Can a linear map from V to W similarly be specified by just a few numbers?

The answer is that it can, at least if V and W are finite dimensional. Suppose that V has a basis $\mathbf{v}_1, \dots, \mathbf{v}_n$, that W has a basis $\mathbf{w}_1, \dots, \mathbf{w}_m$, and that $f : V \rightarrow W$ is the linear map we would like to specify. Since every vector in V can be written in the form $a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n$ and since $f(a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n)$ is always equal to $a_1f(\mathbf{v}_1) + \cdots + a_nf(\mathbf{v}_n)$, once we decide what $f(\mathbf{v}_1), \dots, f(\mathbf{v}_n)$ are we have specified f completely. But each vector $f(\mathbf{v}_j)$ is a linear combination of the basis vectors $\mathbf{w}_1, \dots, \mathbf{w}_m$: that is, it can be written in the form

$$f(\mathbf{v}_j) = a_{1j}\mathbf{w}_1 + \cdots + a_{mj}\mathbf{w}_m.$$

Thus, to specify an individual $f(\mathbf{v}_j)$ needs m numbers, the scalars a_{1j}, \dots, a_{mj} . Since there are n different vectors \mathbf{v}_j , the linear map is determined by the mn numbers a_{ij} , where i runs from 1 to m and j from 1 to n .

These numbers can be written in an array, as follows:

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

An array like this is called a *matrix*. It is important to note that a different choice of basis vectors for V and W would lead to a different matrix, so one often talks of the matrix of f *relative to a given pair of bases* (a basis for V and a basis for W).

Now suppose that f is a linear map from V to W and that g is a linear map from U to V . Then fg stands for the linear map from U to W obtained by doing first g , then f . If the matrices of f and g , relative to certain bases of U , V , and W , are A and B , then what is the matrix of fg ? To work it out, one takes a basis vector \mathbf{u}_k of U and applies to it the function g , obtaining a linear combination $b_{1k}\mathbf{v}_1 + \cdots + b_{nk}\mathbf{v}_n$ of the basis vectors of V . To this linear combination one applies the function f , obtaining a rather complicated linear combination of linear combinations of the basis vectors $\mathbf{w}_1, \dots, \mathbf{w}_m$ of W .

Pursuing this idea, one can calculate that the entry in row i and column j of the matrix P of fg is $a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj}$. This matrix P is called the *product* of A and B and is written AB . If you have not seen this definition then you will find it hard to grasp, but the main point to remember is that there is a way of calculating the matrix for fg from the matrices A , B of f and g , and that this matrix is denoted AB . Matrix multiplication of this kind is associative but not commutative. That is, $A(BC)$ is always equal to $(AB)C$ but AB is not necessarily the same as BA . The associativity follows from the fact that composition of the underlying linear maps is associative: if A , B , and C are the matrices of f , g , and h , respectively, then $A(BC)$ is the matrix of the linear map “do h -then- g , then f ” and $(AB)C$ is the matrix of the linear map “do h , then g -then- f ,” and these are the same linear map.

Let us now confine our attention to *automorphisms* from a vector space V to itself. These are linear maps $f : V \rightarrow V$ that can be inverted; that is, for which there exists a linear map $g : V \rightarrow V$ such that $fg(\mathbf{v}) = g f(\mathbf{v}) = \mathbf{v}$ for every vector \mathbf{v} in V . These we can think of as “symmetries” of the vector space V , and as such they form a group under composition. If V is n dimensional and the scalars come from the field \mathbb{F} , then this group is called $\text{GL}_n(\mathbb{F})$. The letters “G” and “L” stand for “general” and “linear”; some of the most important and difficult problems in mathematics arise when one tries to

understand the structure of the general linear groups (and related groups) for certain interesting fields \mathbb{F} (see REPRESENTATION THEORY [IV.12]).

While matrices are very useful, many interesting linear maps are between infinite-dimensional vector spaces, and we close this section with two examples for the reader who is familiar with elementary calculus. (There will be a brief discussion of calculus later in this article.) For the first, let V be the set of all functions from \mathbb{R} to \mathbb{R} that can be differentiated and let W be the set of *all* functions from \mathbb{R} to \mathbb{R} . These can be made into vector spaces in a simple way: if f and g are functions, then their sum is the function h defined by the formula $h(x) = f(x) + g(x)$, and if a is a real number then af is the function k defined by the formula $k(x) = af(x)$. (So, for example, we could regard the polynomial $x^2 + 3x + 2$ as a linear combination of the functions x^2 , x , and the constant function 1.) Then differentiation is a linear map (from V to W), since the derivative $(af + bg)'$ is $af' + bg'$. This is clearer if we write Df for the derivative of f : then we are saying that $D(af + bg) = aDf + bDg$.

A second example uses integration. Let V be another vector space of functions, and let u be a function of two variables. (The functions involved have to have certain properties for the definition to work, but let us ignore the technicalities.) Then we can define a linear map T on the space V by the formula

$$(Tf)(x) = \int u(x, y)f(y) dy.$$

Definitions like this one can be hard to take in, because they involve holding in one's mind three different levels of complexity. At the bottom we have real numbers, denoted by x and y . In the middle are functions like f , u , and Tf , which turn real numbers (or pairs of them) into real numbers. At the top is another function, T , but the "objects" that it transforms are themselves functions: it turns a function like f into a different function Tf . This is just one example where it is important to think of a function as a single, elementary "thing" rather than as a process of transformation. (See the discussion of functions in THE LANGUAGE AND GRAMMAR OF MATHEMATICS [I.2 §2.2].) Another remark that may help to clarify the definition is that there is a very close analogy between the role of the two-variable function $u(x, y)$ and the role of a matrix a_{ij} (which can itself be thought of as a function of the two integer variables i and j). Functions like u are sometimes called *kernels*. For more about linear maps between infinite-dimensional spaces, see OPERATOR ALGEBRAS [IV.19] and LINEAR OPERATORS [III.52].

4.3 Eigenvalues and Eigenvectors

Let V be a vector space and let $S : V \rightarrow V$ be a linear map from V to itself. An *eigenvector* of S is a nonzero vector v in V such that Sv is proportional to v ; that is, $Sv = \lambda v$ for some scalar λ . The scalar in question is called the *eigenvalue* corresponding to v . This simple pair of definitions is extraordinarily important: it is hard to think of any branch of mathematics where eigenvectors and eigenvalues do not have a major part to play. But what is so interesting about Sv being proportional to v ? A rather vague answer is that in many cases the eigenvectors and eigenvalues associated with a linear map contain all the information one needs about the map, and in a very convenient form. Another answer is that linear maps occur in many different contexts, and questions that arise in those contexts often turn out to be questions about eigenvectors and eigenvalues, as the following two examples illustrate.

First, imagine that you are given a linear map T from a vector space V to itself and want to understand what happens if you perform the map repeatedly. One approach would be to pick a basis of V , work out the corresponding matrix A of T and calculate the powers of A by matrix multiplication. The trouble is that the calculation will be messy and uninformative, and it does not really give much insight into the linear map.

However, it often happens that one can pick a very special basis, consisting only of eigenvectors, and in that case understanding the powers of T becomes easy. Indeed, suppose that the basis vectors are v_1, v_2, \dots, v_n and that each v_i is an eigenvector with corresponding eigenvalue λ_i . That is, suppose that $T(v_i) = \lambda_i v_i$ for every i . If w is any vector in V , then there is exactly one way of writing it in the form $a_1 v_1 + \dots + a_n v_n$, and then

$$T(w) = \lambda_1 a_1 v_1 + \dots + \lambda_n a_n v_n.$$

Roughly speaking, this says that T stretches the part of w in direction v_i by a factor of λ_i . But now it is easy to say what happens if we apply T not just once but m times to w . The result will be

$$T^m(w) = \lambda_1^m a_1 v_1 + \dots + \lambda_n^m a_n v_n.$$

In other words, now the amount by which we stretch in the v_i direction is λ_i^m , and that is all there is to it.

Why should one be interested in doing linear maps over and over again? There are many reasons, but one fairly convincing one is that this sort of calculation is exactly what Google does in order to put Web sites into a useful order. Details can be found in THE MATHEMATICS OF ALGORITHM DESIGN [VII.5].

The second example concerns the interesting property of the EXPONENTIAL FUNCTION [III.25] e^x : that its derivative is the same function. In other words, if $f(x) = e^x$, then $f'(x) = f(x)$. Now differentiation, as we saw earlier, can be thought of as a linear map, and if $f'(x) = f(x)$ then this map leaves the function f unchanged, which says that f is an eigenvector with eigenvalue 1. More generally, if $g(x) = e^{\lambda x}$, then $g'(x) = \lambda e^{\lambda x} = \lambda g(x)$, so g is an eigenvector of the differentiation map, with eigenvalue λ . Many linear differential equations can be thought of as asking for eigenvectors of linear maps defined using differentiation. (Differentiation and differential equations will be discussed in the next section.)

5 Basic Concepts of Mathematical Analysis

Mathematics took a huge leap forward in sophistication with the invention of calculus, and the notion that one can specify a mathematical object indirectly by means of better and better approximations. These ideas form the basis of a broad area of mathematics known as *analysis*, and the purpose of this section is to help the reader who is unfamiliar with them. However, it will not be possible to do full justice to the subject, and what is written here will be hard to understand without at least some prior knowledge of calculus.

5.1 Limits

In our discussion of real numbers (section 1.4) there was a brief discussion of the square root of 2. How do we know that 2 has a square root? One answer is the one given there: that we can calculate its decimal expansion. If we are asked to be more precise, we may well end up saying something like this. The real numbers 1, 1.4, 1.41, 1.414, 1.4142, 1.41421, ..., which have terminating decimal expansions (and are therefore rational) approach another number $x = 1.4142135\dots$. We cannot actually write down x properly because it has an infinite decimal expansion but we can at least explain how its digits are defined: for example, the third digit after the decimal point is a 4 because 1.414 is the largest multiple of 0.001 that squares to less than 2. It follows that the squares of the original numbers, 1, 1.96, 1.9881, 1.999396, 1.99996164, 1.9999899241, ..., approach 2, and this is why we are entitled to say that $x^2 = 2$.

Suppose that we are asked to determine the length of a curve drawn on a piece of paper, and that we are given a ruler to help us. We face a problem: the ruler is straight and the curve is not. One way of tackling the problem is as follows. First, draw a few points $P_0, P_1, P_2, \dots, P_n$ along

the curve, with P_0 at one end and P_n at the other. Next, measure the distance from P_0 to P_1 , the distance from P_1 to P_2 , and so on up to P_n . Finally, add all these distances up. The result will not be an exactly correct answer, but if there are enough points, spaced reasonably evenly, and if the curve does not wiggle too much, then our procedure will give us a good notion of the “approximate length” of the curve. Moreover, it gives us a way to *define* what we mean by the “exact length”: suppose that, as we take more and more points, we find that the approximate lengths, in the sense just defined, approach some number l . Then we say that l is the length of the curve.

In both these examples, there is a number that we reach by means of better and better approximations. I used the word “approach” in both cases, but this is rather vague, and it is important to make it precise. Let a_1, a_2, a_3, \dots be a sequence of real numbers. What does it mean to say that these numbers approach a specified real number l ?

The following two examples are worth bearing in mind. The first is the sequence $\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \dots$. In a sense, the numbers in this sequence approach 2, since each one is closer to 2 than the one before, but it is clear that this is not what we mean. What matters is not so much that we get closer and closer, but that we get *arbitrarily close*, and the only number that is approached in this stronger sense is the obvious “limit,” 1.

A second sequence illustrates this in a different way: $1, 0, \frac{1}{2}, 0, \frac{1}{3}, 0, \frac{1}{4}, 0, \dots$. Here, we would like to say that the numbers approach 0, even though it is not true that each one is closer than the one before. Nevertheless, it is true that eventually the sequence gets as close as you like to 0 and remains at least that close.

This last phrase serves as a definition of the mathematical notion of a *limit*: the limit of the sequence of numbers a_1, a_2, a_3, \dots is l if eventually the sequence gets as close as you like to l and remains that close. However, in order to meet the standards of precision demanded by mathematics, we need to know how to translate English words like “eventually” into mathematics, and for this we need QUANTIFIERS [I.2 §3.2].

Suppose δ is a positive number (which one usually imagines as small). Let us say that a_n is δ -close to l if $|a_n - l|$, the difference between a_n and l , is less than δ . What would it mean to say that eventually the sequence gets δ -close to l and stays there? It means that from some point onwards, all the a_n are δ -close to l . And what is the meaning of “from some point onwards”? It is that there is some number N (the point in question) with the property that a_n is δ -close to l from N onwards—that is,

for every n that is greater than or equal to N . In symbols:

$$\exists N \quad \forall n \geq N \quad a_n \text{ is } \delta\text{-close to } l.$$

It remains to capture the idea of “as close as you like.” What this means is that the above sentence is true for any δ you might wish to specify. In symbols:

$$\forall \delta > 0 \quad \exists N \quad \forall n \geq N \quad a_n \text{ is } \delta\text{-close to } l.$$

Finally, let us stop using the nonstandard phrase “ δ -close”:

$$\forall \delta > 0 \quad \exists N \quad \forall n \geq N \quad |a_n - l| < \delta.$$

This sentence is not particularly easy to understand. Unfortunately (and interestingly in the light of the discussion in [I.2 §4]), using a less symbolic language does not necessarily make things much easier: “Whatever positive δ you choose, there is some number N such that for all bigger numbers n the difference between a_n and l is less than δ .”

The notion of limit applies much more generally than just to real numbers. If you have any collection of mathematical objects and can say what you mean by the distance between any two of those objects, then you can talk of a sequence of those objects having a limit. Two objects are now called δ -close if the *distance* between them is less than δ , rather than the difference. (The idea of distance is discussed further in METRIC SPACES [III.58].) For example, a sequence of points in space can have a limit, as can a sequence of functions. (In the second case it is less obvious how to define distance—there are many natural ways to do it.) A further example comes in the theory of fractals (see DYNAMICS [IV.15]): the very complicated shapes that appear there are best defined as limits of simpler ones.

Other ways of saying that the limit of the sequence a_1, a_2, \dots is l are to say that a_n *converges to* l or that it *tends to* l . One sometimes says that this happens *as n tends to infinity*. Any sequence that has a limit is called *convergent*. If a_n converges to l then one often writes $a_n \rightarrow l$.

5.2 Continuity

Suppose you want to know the approximate value of π^2 . Perhaps the easiest thing to do is to press a π button on a calculator, which displays 3.1415927, and then an x^2 button, after which it displays 9.8696044. Of course, one knows that the calculator has not actually squared π : instead it has squared the number 3.1415927. (If it is a good one, then it may have secretly used a few more digits of π without displaying them, but not infinitely many.) Why does it not matter that the calculator has squared the wrong number?

A first answer is that it was only an *approximate* value of π^2 that was required. But that is not quite a complete explanation: how do we know that if x is a good approximation to π then x^2 is a good approximation to π^2 ? Here is how one might show this. If x is a good approximation to π , then we can write $x = \pi + \delta$ for some very small number δ (which could be negative). Then $x^2 = \pi^2 + 2\delta\pi + \delta^2$. Since δ is small, so is $2\delta\pi + \delta^2$, so x^2 is indeed a good approximation to π^2 .

What makes the above reasoning work is that the function that takes a number x to its square is *continuous*. Roughly speaking, this means that if two numbers are close, then so are their squares.

To be more precise about this, let us return to the calculation of π^2 , and imagine that we wish to work it out to a much greater accuracy—so that the first hundred digits after the decimal point are correct, for example. A calculator will not be much help, but what we might do is find a list of the digits of π (on the Internet you can find sites that tell you at least the first fifty million), use this to define a new x that is a much better approximation to π , and then calculate the new x^2 by getting a computer to do the necessary long multiplication.

How close do we need x to be to π for x^2 to be within 10^{-100} of π^2 ? To answer this, we can use our earlier argument. Let $x = \pi + \delta$ again. Then $x^2 - \pi^2 = 2\delta\pi + \delta^2$, and an easy calculation shows that this has modulus less than 10^{-100} if δ has modulus less than 10^{-101} . So we will be all right if we take the first 101 digits of π after the decimal point.

More generally, *however* accurate we wish our estimate of π^2 to be, we can achieve this accuracy if we are prepared to make x a sufficiently good approximation to π . In mathematical parlance, the function $f(x) = x^2$ is *continuous at* π .

Let us try to say this more symbolically. The statement “ $x^2 = \pi^2$ to within an accuracy of ϵ ” means that $|x^2 - \pi^2| < \epsilon$. To capture the phrase “however accurate,” we need this to be true for every positive ϵ , so we should start by saying $\forall \epsilon > 0$. Now let us think about the words “if we are prepared to make x a sufficiently good approximation to π .” The thought behind them is that there is some $\delta > 0$ for which the approximation is guaranteed to be accurate to within ϵ as long as x is within δ of π . That is, there exists a $\delta > 0$ such that if $|x - \pi| < \delta$ then it is guaranteed that $|x^2 - \pi^2| < \epsilon$. Putting everything together, we end up with the following symbolic sentence:

$$\forall \epsilon > 0 \quad \exists \delta > 0 \quad (|x - \pi| < \delta \Rightarrow |x^2 - \pi^2| < \epsilon).$$

To put that in words: “Given any positive number ϵ there is a positive number δ such that if $|x - \pi|$ is less than δ

then $|x^2 - \pi^2|$ is less than ϵ ." Earlier, we found a δ that worked when ϵ was chosen to be 10^{-100} ; it was 10^{-101} .

What we have just shown is that the function $f(x) = x^2$ is continuous at the point $x = \pi$. Now let us generalize this idea: let f be any function and let a be any real number. We say that f is *continuous at a* if

$$\forall \epsilon > 0 \quad \exists \delta > 0 \quad (|x - a| < \delta \Rightarrow |f(x) - f(a)| < \epsilon).$$

This says that however accurate you wish $f(x)$ to be as an estimate for $f(a)$, you can achieve this accuracy if you are prepared to make x a sufficiently good approximation to a . The function f is said to be *continuous* if it is continuous at every a . Roughly speaking, what this means is that f has no "sudden jumps." (It also rules out certain kinds of very rapid oscillations that would also make accurate estimates difficult.)

As with limits, the idea of continuity applies in much more general contexts, and for the same reason. Let f be a function from a set X to a set Y (see THE LANGUAGE AND GRAMMAR OF MATHEMATICS [I.2 §2.2]), and suppose that we have two notions of distance, one for elements of X and the other for elements of Y . Using the expression $d(x, a)$ to denote the distance between x and a , and similarly for $d(f(x), f(a))$, one says that f is *continuous at a* if

$$\forall \epsilon > 0 \quad \exists \delta > 0 \quad (d(x, a) < \delta \Rightarrow d(f(x), f(a)) < \epsilon)$$

and that f is *continuous* if it is continuous at every a in X . In other words, we replace differences such as $|x - a|$ by distances such as $d(x, a)$.

Continuous functions, like HOMOMORPHISMS (see section 4.1 above), can be regarded as preserving a certain sort of structure. It can be shown that a function f is continuous if and only if, whenever $a_n \rightarrow x$, we also have $f(a_n) \rightarrow f(x)$. That is, continuous functions are functions that preserve the structure provided by convergent sequences and their limits.

5.3 Differentiation

The derivative of a function f at a value a is usually presented as a number that measures the rate of change of $f(x)$ as x passes through a . The purpose of this section is to promote a slightly different way of regarding it, one that is more general and that opens the door to much of modern mathematics. This is the idea of differentiation as *linear approximation*.

Intuitively speaking, to say that $f'(a) = m$ is to say that if one looks through a very powerful microscope at the graph of f in a tiny region that includes the point $(a, f(a))$, then what one sees is almost exactly a straight line of gradient m . In other words, in a sufficiently small

neighborhood of the point a , the function f is approximately linear. We can even write down a formula for the linear function g that approximates f :

$$g(x) = f(a) + m(x - a).$$

This is the equation of the straight line of gradient m that passes through the point $(a, f(a))$. Another way of writing it, which is a little clearer, is

$$g(a + h) = f(a) + mh,$$

and to say that g approximates f in a small neighborhood of a is to say that $f(a + h)$ is *approximately* equal to $f(a) + mh$ when h is small.

One must be a little careful here: after all, if f does not jump suddenly, then, when h is small, $f(a + h)$ will be close to $f(a)$ and mh will be small, so $f(a + h)$ is approximately equal to $f(a) + mh$. This line of reasoning seems to work regardless of the value of m , and yet we wanted there to be something special about the choice $m = f'(a)$. What singles out that particular value is that $f(a + h)$ is not just close to $f(a) + mh$, but the difference $\epsilon(h) = f(a + h) - f(a) - mh$ is small *compared with h* . That is, $\epsilon(h)/h \rightarrow 0$ as $h \rightarrow 0$. (This is a slightly more general notion of limit than that discussed in section 5.1, but can be recovered from it: it is equivalent to saying that if you choose any sequence h_1, h_2, \dots such that $h_n \rightarrow 0$, then $\epsilon(h_n)/h_n \rightarrow 0$ as well.)

The reason these ideas can be generalized is that the notion of a linear map is much more general than simply a function from \mathbb{R} to \mathbb{R} of the form $g(x) = mx + c$. Many functions that arise naturally in mathematics—and also in science, engineering, economics, and many other areas—are functions of *several variables*, and can therefore be regarded as functions defined on a vector space of dimension greater than 1. As soon as we look at them this way, we can ask ourselves whether, in a small neighborhood of a point, they can be approximated by linear maps. It is very useful if they can: a general function can behave in very complicated ways, but if it can be approximated by a linear function, then at least in small regions of n -dimensional space its behavior is much easier to understand. In this situation one can use the machinery of linear algebra and matrices, which leads to calculations that are feasible, especially if one has the help of a computer.

Imagine, for instance, a meteorologist interested in how the direction and speed of the wind changes as one looks at different parts of some three-dimensional region above Earth's surface. Wind behaves in complicated, chaotic ways, but to get some sort of handle on this behavior one can describe it as follows. To each

point (x, y, z) in the region (think of x and y as horizontal coordinates and z as a vertical one) one can associate a vector (u, v, w) representing the velocity of the wind at that point: u , v , and w are the components of the velocity in the x -, y -, and z -directions.

Now let us change the point (x, y, z) very slightly by choosing three small numbers h, k , and l and looking at $(x + h, y + k, z + l)$. At this new point, we would expect the wind vector to be slightly different as well, so let us write it $(u + p, v + q, w + r)$. How does the small change (p, q, r) in the wind vector depend on the small change (h, k, l) in the position vector? Provided the wind is not too turbulent and h, k , and l are small enough, we expect the dependence to be roughly linear: that is how nature seems to work. In other words, we expect there to be some linear map T such that (p, q, r) is roughly $T(h, k, l)$ when h, k , and l are small. Notice that each of p, q , and r depends on each of h, k , and l , so nine numbers will be needed in order to specify this linear map. In fact, we can express it in matrix form:

$$\begin{pmatrix} p \\ q \\ r \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} h \\ k \\ l \end{pmatrix}.$$

The matrix entries a_{ij} express individual dependencies. For example, if x and z are held fixed, then we are setting $h = l = 0$, from which it follows that the rate of change u as just y varies is given by the entry a_{12} . That is, a_{12} is the *partial derivative* $\partial u / \partial y$ at the point (x, y, z) .

This tells us how to calculate the matrix, but from the conceptual point of view it is easier to use vector notation. Write \mathbf{x} for (x, y, z) , $\mathbf{u}(\mathbf{x})$ for (u, v, w) , \mathbf{h} for (h, k, l) , and \mathbf{p} for (p, q, r) . Then what we are saying is that

$$\mathbf{p} = T(\mathbf{h}) + \boldsymbol{\epsilon}(\mathbf{h})$$

for some vector $\boldsymbol{\epsilon}(\mathbf{h})$ that is small relative to \mathbf{h} . Alternatively, we can write

$$\mathbf{u}(\mathbf{x} + \mathbf{h}) = \mathbf{u}(\mathbf{x}) + T(\mathbf{h}) + \boldsymbol{\epsilon}(\mathbf{h}),$$

a formula that is closely analogous to our earlier formula $g(\mathbf{x} + \mathbf{h}) = g(\mathbf{x}) + \mathbf{m}\mathbf{h} + \boldsymbol{\epsilon}(\mathbf{h})$. This tells us that if we add a small vector \mathbf{h} to \mathbf{x} , then $\mathbf{u}(\mathbf{x})$ will change by roughly $T(\mathbf{h})$.

5.4 Partial Differential Equations

Partial differential equations are of immense importance in physics, and have inspired a vast amount of mathematical research. Three basic examples will be discussed here, as an introduction to more advanced articles later in the volume (see, in particular, PARTIAL DIFFERENTIAL EQUATIONS [IV.16]).

The first is the *heat equation*, which, as its name suggests, describes the way the distribution of heat in a physical medium changes with time:

$$\frac{\partial T}{\partial t} = \kappa \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} \right).$$

Here, $T(x, y, z, t)$ is a function that specifies the temperature at the point (x, y, z) at time t .

It is one thing to read an equation like this and understand the symbols that make it up, but quite another to see what it really means. However, it is important to do so, since of the many expressions one could write down that involve partial derivatives, only a minority are of much significance, and these tend to be the ones that have interesting interpretations. So let us try to interpret the expressions involved in the heat equation.

The left-hand side, $\partial T / \partial t$, is quite simple. It is the rate of change of the temperature $T(x, y, z, t)$ when the spatial coordinates x, y , and z are kept fixed and t varies. In other words, it tells us how fast the point (x, y, z) is heating up or cooling down at time t . What would we expect this to depend on? Well, heat takes time to travel through a medium, so although the temperature at some distant point (x', y', z') will eventually affect the temperature at (x, y, z) , the way the temperature is changing *right now* (that is, at time t) will be affected only by the temperatures of points very close to (x, y, z) : if points in the immediate neighborhood of (x, y, z) are hotter, on average, than (x, y, z) itself, then we expect the temperature at (x, y, z) to be increasing, and if they are colder then we expect it to be decreasing.

The expression in brackets on the right-hand side appears so often that it has its own shorthand. The symbol Δ , defined by

$$\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2},$$

is known as the *Laplacian*. What information does Δf give us about a function f ? The answer is that it captures the idea in the last paragraph: it tells us how the value of f at (x, y, z) compares with the average value of f in a small neighborhood of (x, y, z) , or, more precisely, with the limit of the average value in a neighborhood of (x, y, z) as the size of that neighborhood shrinks to zero.

This is not immediately obvious from the formula, but the following (not wholly rigorous) argument in one dimension gives a clue about why second derivatives should be involved. Let f be a function that takes real numbers to real numbers. Then to obtain a good approximation to the second derivative of f at a point x , one can look at the expression $(f'(x) - f'(x - h))/h$

for some small h . (If one substitutes $-h$ for h in the above expression, one obtains the more usual formula, but this one is more convenient here.) The derivatives $f'(x)$ and $f'(x-h)$ can themselves be approximated by $(f(x+h) - f(x))/h$ and $(f(x) - f(x-h))/h$, respectively, and if we substitute these approximations into the earlier expression, then we obtain

$$\frac{1}{h} \left(\frac{f(x+h) - f(x)}{h} - \frac{f(x) - f(x-h)}{h} \right),$$

which equals $(f(x+h) - 2f(x) + f(x-h))/h^2$. Dividing the top of this last fraction by 2, we obtain $\frac{1}{2}(f(x+h) + f(x-h)) - f(x)$: that is, the difference between the value of f at x and the average value of f at the two surrounding points $x+h$ and $x-h$.

In other words, the second derivative conveys just the idea we want—a comparison between the value at x and the average value near x . It is worth noting that if f is linear, then the average of $f(x-h)$ and $f(x+h)$ will be equal to $f(x)$, which fits with the familiar fact that the second derivative of a linear function f is zero.

Just as, when defining the first derivative, we have to divide the difference $f(x+h) - f(x)$ by h so that it is not automatically tiny, so with the second derivative it is appropriate to divide by h^2 . (This is appropriate, since, whereas the first derivative concerns linear approximations, the second derivative concerns *quadratic* ones: the best quadratic approximation for a function f near a value x is $f(x+h) = f(x) + hf'(x) + \frac{1}{2}h^2f''(x)$, an approximation that one can check is exact if f was a quadratic function to start with.)

It is possible to pursue thoughts of this kind and show that if f is a function of three variables then the value of Δf at (x, y, z) does indeed tell us how the value of f at (x, y, z) compares with the average values of f at points nearby. (There is nothing special about the number 3 here—the ideas can easily be generalized to functions of any number of variables.) All that is left to discuss in the heat equation is the parameter κ . This measures the *conductivity* of the medium. If κ is small, then the medium does not conduct heat very well and ΔT has less of an effect on the rate of change of the temperature; if it is large then heat is conducted better and the effect is greater.

A second equation of great importance is the *Laplace equation*, $\Delta f = 0$. Intuitively speaking, this says of a function f that its value at a point (x, y, z) is always equal to the average value at the immediately surrounding points. If f is a function of just one variable x , this says that the second derivative of f is zero, which implies that f is of the form $ax + b$. However, for two or more variables, a function has more flexibility—it can lie

above the tangent lines in some directions and below it in others. As a result, one can impose a variety of boundary conditions on f (that is, specifications of the values f takes on the boundaries of certain regions), and there is a much wider and more interesting class of solutions.

A third fundamental equation is the *wave equation*. In its one-dimensional formulation it describes the motion of a vibrating string that connects two points A and B. Suppose that the height of the string at distance x from A and at time t is written $h(x, t)$. Then the wave equation says that

$$\frac{1}{v^2} \frac{\partial^2 h}{\partial t^2} = \frac{\partial^2 h}{\partial x^2}.$$

Ignoring the constant $1/v^2$ for a moment, the left-hand side of this equation represents the acceleration (in a vertical direction) of the piece of string at distance x from A. This should be proportional to the force acting on it. What will govern this force? Well, suppose for a moment that the portion of string containing x were absolutely straight. Then the pull of the string on the left of x would exactly cancel out the pull on the right and the net force would be zero. So, once again, what matters is how the height at x compares with the average height on either side: if the string lies above the tangent line at x , then there will be an upwards force, and if it lies below, then there will be a downwards one. This is why the second derivative appears on the right-hand side once again. How much force results from this second derivative depends on factors such as the density and tautness of the string, which is where the constant comes in. Since h and x are both distances, v^2 has dimensions of (distance/time)², which means that v represents a speed, which is, in fact, the speed of propagation of the wave.

Similar considerations yield the three-dimensional wave equation, which is, as one might now expect,

$$\frac{1}{v^2} \frac{\partial^2 h}{\partial t^2} = \frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2} + \frac{\partial^2 h}{\partial z^2},$$

or, more concisely,

$$\frac{1}{v^2} \frac{\partial^2 h}{\partial t^2} = \Delta h.$$

One can be more concise still and write this equation as $\square^2 h = 0$, where $\square^2 h$ is shorthand for

$$\Delta h - \frac{1}{v^2} \frac{\partial^2 h}{\partial t^2}.$$

The operation \square^2 is called the *d'Alembertian*, after D'ALEMBERT [VI.19], who was the first to formulate the wave equation.

5.5 Integration

Suppose that a car drives down a long straight road for one minute, and that you are told where it starts and what its speed is during that minute. How can you work out how far it has gone? If it travels at the same speed for the whole minute then the problem is very simple indeed—for example, if that speed is thirty miles per hour then we can divide by sixty and see that it has gone half a mile—but the problem becomes more interesting if the speed varies. Then, instead of trying to give an exact answer, one can use the following technique to approximate it. First, write down the speed of the car at the beginning of each of the sixty seconds that it is traveling. Next, for each of those seconds, do a simple calculation to see how far the car would have gone during that second if the speed had remained exactly as it was at the beginning of the second. Finally, add up all these distances. Since one second is a short time, the speed will not change very much during any one second, so this procedure gives quite an accurate answer. Moreover, if you are not satisfied with this accuracy, then you can improve it by using intervals that are shorter than a second.

If you have done a first course in calculus, then you may well have solved such problems in a completely different way. In a typical question, one is given an explicit formula for the speed at time t —something like $at + u$, for example—and in order to work out how far the car has gone one “integrates” this function to obtain the formula $\frac{1}{2}at^2 + ut$ for the distance traveled at time t . Here, integration simply means the opposite of differentiation: to find the integral of a function f is to find a function g such that $g'(t) = f(t)$. This makes sense, because if $g(t)$ is the distance traveled and $f(t)$ is the speed, then $f(t)$ is indeed the rate of change of $g(t)$.

However, antidifferentiation is not the *definition* of integration. To see why not, consider the following question: what is the distance traveled if the speed at time t is e^{-t^2} . It is known that there is no nice function (which means, roughly speaking, a function built up out of standard ones such as polynomials, exponentials, logarithms, and trigonometric functions) with e^{-t^2} as its derivative, yet the question still makes good sense and has a definite answer. (It is possible that you have heard of a function $\Phi(t)$ that differentiates to $e^{-t^2/2}$, from which it follows that $\Phi(t\sqrt{2})/\sqrt{2}$ differentiates to e^{-t^2} . However, this does not remove the difficulty, since $\Phi(t)$ is defined as the integral of $e^{-t^2/2}$.)

In order to define integration in situations like this where antidifferentiation runs into difficulties, we must

fall back on messy approximations of the kind discussed earlier. A formal definition along such lines was given by RIEMANN [VI.48] in the mid nineteenth century. To see what Riemann’s basic idea is, and to see also that integration, like differentiation, is a procedure that can usefully be applied to functions of more than one variable, let us look at another physical problem.

Suppose that you have a lump of impure rock and wish to calculate its mass from its density. Suppose also that this density is not constant but varies rather irregularly through the rock. Perhaps there are even holes inside, so that the density is zero in places. What should you do?

Riemann’s approach would be this. First, you enclose the rock in a cuboid. For each point (x, y, z) in this cuboid there is then an associated density $d(x, y, z)$ (which will be zero if (x, y, z) lies outside the rock or inside a hole). Second, you divide the cuboid into a large number of smaller cuboids. Third, in each of the small cuboids you look for the point of lowest density (if any point in the cuboid is not in the rock, then this density will be zero) and the point of highest density. Let C be one of the small cuboids and suppose that the lowest and highest densities in C are a and b , respectively, and that the volume of C is V . Then the mass of the part of the rock that lies in C must lie between aV and bV . Fourth, add up all the numbers aV that are obtained in this way, and then add up all the numbers bV . If the totals are M_1 and M_2 , respectively, then the total mass of rock has to lie between M_1 and M_2 . Finally, repeat this calculation for subdivisions into smaller and smaller cuboids. As you do this, the resulting numbers M_1 and M_2 will become closer and closer to each other, and you will have better and better approximations to the mass of the rock.

Similarly, his approach to the problem about the car would be to divide the minute up into small intervals and look at the minimum and maximum speeds during those intervals. This would enable him to say for each interval that the car had traveled a distance of at least a and at most b . Adding up these sets of numbers, he could then say that over the full minute the car must have traveled a distance of at least D_1 (the sum of the as) and at most D_2 (the sum of the bs).

For both these problems we had a function (density/speed) defined on a set (the cuboid/a minute of time) and in a certain sense we wanted to work out the “total amount” of the function. We did so by dividing the set into small parts and doing simple calculations in those parts to obtain approximations to this amount from below and above. This process is what is known

PUP: to solve antecedent problem spotted by proofreader in the next sentence, Tim rewrote this one. OK?

as (Riemann) *integration*. The following notation is common: if S is the set and f is the function, then the total amount of f in S , known as the *integral*, is written $\int_S f(x) dx$. Here, x denotes a typical element of S . If, as in the density example, the elements of S are points (x, y, z) , then vector notation such as $\int_S f(\mathbf{x}) d\mathbf{x}$ can be used, though often it is not and the reader is left to deduce from the context that an ordinary “ x ” denotes a vector rather than a real number.

We have been at pains to distinguish integration from antidifferentiation, but a famous theorem, known as the *fundamental theorem of calculus*, asserts that the two procedures do, in fact, give the same answer, at least when the function in question has certain continuity properties that all “sensible” functions have. So it is usually legitimate to regard integration as the opposite of differentiation. More precisely, if f is continuous and $F(x)$ is defined to be $\int_a^x f(t) dt$ for some a , then F can be differentiated and $F'(x) = f(x)$. That is, if you integrate a continuous function and differentiate it again, you get back to where you started. Going the other way around, if F has a continuous derivative f and $a < b$, then $\int_a^b f(t) dt = F(b) - F(a)$. This almost says that if you differentiate F and then integrate it again, you get back to F . Actually, you have to choose an arbitrary number a and what you get is the function F with the constant $F(a)$ subtracted.

To give an idea of the sort of exceptions that arise if one does not assume continuity, consider the so-called *Heaviside step function* $H(x)$, which is 0 when $x < 0$ and 1 when $x \geq 0$. This function has a jump at 0 and is therefore not continuous. The integral $J(x)$ of this function is 0 when $x < 0$ and x when $x \geq 0$, and for almost all values of x we have $J'(x) = H(x)$. However, the gradient of J suddenly changes at 0, so J is not differentiable there and one cannot say that $J'(0) = H(0) = 1$.

5.6 Holomorphic Functions

One of the jewels in the crown of mathematics is *complex analysis*, which is the study of differentiable functions that take complex numbers to complex numbers. Functions of this kind are called *holomorphic*.

At first, there seems to be nothing special about such functions, since the definition of a derivative in this context is no different from the definition for functions of a real variable: if f is a function then the derivative $f'(z)$ at a complex number z is defined to be the limit as h tends to zero of $(f(z+h) - f(z))/h$. However, if we look at this definition in a slightly different way (one which we saw in section 5.3), we find that it is not altogether easy for a complex function to be differentiable.

Recall from that section that differentiation means *linear approximation*. In the case of a complex function, this means that we would like to approximate it by functions of the form $g(w) = \lambda w + \mu$, where λ and μ are complex numbers. (The approximation near z will be $g(w) = f(z) + f'(z)(w - z)$, which gives $\lambda = f'(z)$ and $\mu = f(z) - zf'(z)$.)

Let us regard this situation geometrically. If $\lambda \neq 0$ then the effect of multiplying by λ is to expand z by some factor r and to rotate it by some angle θ . This means that many transformations of the plane that we would ordinarily consider to be linear, such as reflections, shears, or stretches, are ruled out. We need two real numbers to specify λ (whether we write it in the form $a + bi$ or $re^{i\theta}$), but to specify a general linear transformation of the plane takes four (see the discussion of matrices in section 4.2). This reduction in the number of degrees of freedom is expressed by a pair of differential equations called the *Cauchy-Riemann equations*. Instead of writing $f(z)$ let us write $u(x + iy) + iv(x + iy)$, where x and y are the real and imaginary parts of z and $u(x + iy)$ and $v(x + iy)$ are the real and imaginary parts of $f(x + iy)$. Then the linear approximation to f near z has the matrix

$$\begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix}.$$

The matrix of an expansion and rotation always has the form $\begin{pmatrix} a & b \\ -b & a \end{pmatrix}$, from which we deduce that

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \quad \text{and} \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}.$$

These are the Cauchy-Riemann equations. One consequence of these equations is that

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \frac{\partial^2 v}{\partial x \partial y} - \frac{\partial^2 v}{\partial y \partial x} = 0.$$

(It is not obvious that the necessary conditions hold for the symmetry of the mixed partial derivatives, but when f is holomorphic they do.) Therefore, u satisfies the Laplace equation (which was discussed in section 5.4). A similar argument shows that v does as well.

These facts begin to suggest that complex differentiability is a much stronger condition than real differentiability and that we should expect holomorphic functions to have interesting properties. For the remainder of this subsection, let us look at a few of the remarkable properties that they do indeed have.

The first is related to the fundamental theorem of calculus (discussed in the previous subsection). Suppose that F is a holomorphic function and we are given its derivative f and the value of $F(u)$ for some complex

PUP: change to cross-reference OK here?

number u . How can we reconstruct F ? An approximate method is as follows. Let w be another complex number and let us try to work out $F(w)$. We take a sequence of points z_0, z_1, \dots, z_n with $z_0 = u$ and $z_n = z$, and with the differences $|z_1 - z_0|, |z_2 - z_1|, \dots, |z_n - z_{n-1}|$ all small. We can then approximate $F(z_{i+1}) - F(z_i)$ by $(z_{i+1} - z_i)f(z_i)$. It follows that $F(w) - F(u)$, which equals $F(z_n) - F(z_0)$, is approximated by the sum of all the $(z_{i+1} - z_i)f(z_i)$. (Since we have added together many small errors, it is not obvious that this approximation is a good one, but it turns out that it is.) We can imagine a number z that starts at u and follows a path P to w by jumping from one z_i to another in small steps of $\delta z = z_{i+1} - z_i$. In the limit as n goes to infinity and the steps δz go to zero we obtain a so-called *path integral*, which is denoted $\int_P f(z) dz$.

The above argument has the consequence that if the path P begins and ends at the same point u , then the path integral $\int_P f(z) dz$ is zero. Equivalently, if two paths P_1 and P_2 have the same starting point u and the same endpoint w , then the path integrals $\int_{P_1} f(z) dz$ and $\int_{P_2} f(z) dz$ are the same, since they both give the value $F(w) - F(u)$.

Of course, in order to establish this, we made the big assumption that f was the derivative of a function F . Cauchy's theorem says that the same conclusion is true if f is holomorphic. That is, rather than requiring f to be the derivative of another function, it asks for f itself to have a derivative. If that is the case, then any path integral of f depends only on where the path begins and ends. What is more, these path integrals can be used to define a function F that differentiates to f , so a function with a derivative automatically has an antiderivative.

It is not necessary for the function f to be defined on the whole of \mathbb{C} for Cauchy's theorem to be valid: everything remains true if we restrict attention to a *simply connected domain*, which means an open set with no holes in it. If there are holes, then two path integrals may differ if the paths go around the holes in different ways. Thus, path integrals have a close connection with the *topology* of subsets of the plane, an observation that has many ramifications throughout modern geometry. For more on topology, see section 6.4 of this article and ALGEBRAIC TOPOLOGY [IV.10].

A very surprising fact, which can be deduced from Cauchy's theorem, is that if f is holomorphic then it can be differentiated twice. (This is completely untrue of real-valued functions: consider, for example, the function f where $f(x) = 0$ when $x < 0$ and $f(x) = x^2$ when $x \geq 0$.) It follows that f' is holomorphic, so it too can be differentiated twice. Continuing, one finds that f can

be differentiated any number of times. Thus, for complex functions differentiability implies infinite differentiability. (This property is what is used to establish the symmetry, and even the existence, of the mixed partial derivatives mentioned earlier.)

A closely related fact is that wherever a holomorphic function is defined it can be expanded in a power series. That is, if f is defined and differentiable everywhere on an open disk of radius R about w , then it will be given by a formula of the form

$$f(z) = \sum_{n=0}^{\infty} a_n (z - w)^n$$

valid everywhere in that disk. This is called the *Taylor expansion* of f .

Another fundamental property of holomorphic functions, one that shows just how “rigid” they are, is that their entire behavior is determined just by what they do in a small region. That is, if f and g are holomorphic and they take the same values in some tiny disk, then they must take the same values everywhere. This remarkable fact allows a process of *analytic continuation*. If it is difficult to define a holomorphic function f everywhere you want it defined, then you can simply define it in some small region and say that elsewhere it takes the only possible values that are consistent with the ones that you have just specified. This is how the famous RIEMANN ZETA FUNCTION [IV.4 §3] is conventionally defined.

6 What Is Geometry?

It is not easy to do justice to geometry in this article because the fundamental concepts of the subject are either too simple to need explaining—for example, there is no need to say here what a circle, line, or plane is—or sufficiently advanced that they are better discussed in parts III and IV of the book. However, if you have not met the advanced concepts and have no idea what modern geometry is like, then you will get much more out of this book if you understand two basic ideas: the relationship between geometry and symmetry, and the notion of a manifold. These ideas will occupy us for the rest of the article.

6.1 Geometry and Symmetry Groups

Broadly speaking, geometry is the part of mathematics that involves the sort of language that one would conventionally regard as geometrical, with words such as “point,” “line,” “plane,” “space,” “curve,” “sphere,” “cube,” “distance,” and “angle” playing a prominent role. However, there is a more sophisticated view, first

PUP: proofreader wanted a comma here but Tim would strongly prefer not to insert one. OK to keep it as it is I presume?

advocated by KLEIN [VI.56], which regards *transformations* as the true subject matter of geometry. So, to the above list one should add words like “reflection,” “rotation,” “translation,” “stretch,” “shear,” and “projection,” together with slightly more nebulous concepts such as “angle-preserving map” or “continuous deformation.”

As was discussed in section 2.1, transformations go hand in hand with groups, and for this reason there is an intimate connection between geometry and group theory. Indeed, given any group of transformations, there is a corresponding notion of geometry, in which one studies the phenomena that are unaffected by transformations in that group. In particular, two shapes are regarded as *equivalent* if one can be turned into the other by means of one of the transformations in the group. Different groups will of course lead to different notions of equivalence, and for this reason mathematicians frequently talk about *geometries*, rather than about a single monolithic subject called geometry. This subsection contains brief descriptions of some of the most important geometries and their associated groups of transformations.

6.2 Euclidean Geometry

Euclidean geometry is what most people would think of as “ordinary” geometry, and, not surprisingly given its name, it includes the basic theorems of Greek geometry that were the staple of geometers for thousands of years. For example, the theorem that the three angles of a triangle add up to 180° belongs to Euclidean geometry.

To understand Euclidean geometry from a transformational viewpoint, we need to say how many dimensions we are working in, and we must of course specify a group of transformations. The appropriate group is the group of *rigid* transformations. These can be thought of in two different ways. One is that they are the transformations of the plane, or of space, or more generally of \mathbb{R}^n for some n , that *preserve distance*. That is, T is a rigid transformation if, given any two points x and y , the distance between Tx and Ty is always the same as the distance between x and y . (In dimensions greater than 3, distance is defined in a way that naturally generalizes the Pythagorean formula. See METRIC SPACES [III.58] for more details.)

It turns out that every such transformation can be realized as a combination of rotations, reflections, and translations, and this gives us a more concrete way to think about the group. Euclidean geometry, in other words, is the study of concepts that do not change when you rotate, reflect, or translate, and these include points,

lines, planes, circles, spheres, distance, angle, length, area, and volume. The rotations of \mathbb{R}^n form an important group, the *special orthogonal group*, known as $SO(n)$. The larger *orthogonal group* $O(n)$ includes reflections as well. (It is not quite obvious how to define a “rotation” of n -dimensional space, but it is not too hard to do. An *orthogonal map* of \mathbb{R}^n is a linear map T that preserves distances, in the sense that $d(Tx, Ty)$ is always the same as $d(x, y)$. It is a *rotation* if its DETERMINANT [III.15] is 1. The only other possibility for the determinant of a distance-preserving map is -1 . Such maps are like reflections in that they turn space “inside out.”)

6.3 Affine Geometry

There are many linear maps besides rotations and reflections. What happens if we enlarge our group from $SO(n)$ or $O(n)$ to include as many of them as possible? For a transformation to be part of a group it must be *invertible* and not all linear maps are, so the natural group to look at is the group $GL_n(\mathbb{R})$ of all invertible linear transformations of \mathbb{R}^n , a group that we first met in section 4.2. These maps all leave the origin fixed, but if we want we can incorporate translations and consider a larger group that consists of all transformations of the form $x \mapsto Tx + b$, where b is a fixed vector and T is an invertible linear map. The resulting geometry is called *affine* geometry.

Since linear maps include stretches and shears, they preserve neither distance nor angle, so these are not concepts of affine geometry. However, points, lines, and planes remain as points, lines, and planes after an invertible linear map and a translation, so these concepts do belong to affine geometry. Another affine concept is that of two lines being parallel. (That is, although angles in general are not preserved by linear maps, angles of zero are.) This means that although there is no such thing as a square or a rectangle in affine geometry, one can still talk about a parallelogram. Similarly, one cannot talk of circles but one can talk of ellipses, since a linear map transformation of an ellipse is another ellipse (provided that one regards a circle as a special kind of ellipse).

6.4 Topology

The idea that the geometry associated with a group of transformations “studies the concepts that are preserved by all the transformations” can be made more precise using the notion of EQUIVALENCE RELATIONS [I.2 §2.3]. Indeed, let G be a group of transformations of \mathbb{R}^n . We might think of a d -dimensional “shape” as being a subset S of \mathbb{R}^n , but if we are doing G -geometry, then

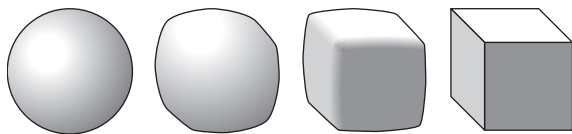


Figure 1 A sphere morphing into a cube.

we do not want to distinguish between a set S and any other set we can obtain from it using a transformation in G . So in that case we say that the two shapes are *equivalent*. For example, two shapes are equivalent in Euclidean geometry if and only if they are congruent in the usual sense, whereas in two-dimensional affine geometry all parallelograms are equivalent, as are all ellipses. One can think of the basic objects of G -geometry as *equivalence classes* of shapes rather than the shapes themselves.

Topology can be thought of as the geometry that arises when we use a particularly generous notion of equivalence, saying that two shapes are equivalent, or *homeomorphic*, to use the technical term, if each can be “continuously deformed” into the other. For example, a sphere and a cube are equivalent in this sense, as figure 1 illustrates.

Because there are very many continuous deformations, it is quite hard to prove that two shapes are *not* equivalent in this sense. For example, it may seem obvious that a sphere (this means the surface of a ball rather than the solid ball) cannot be continuously deformed into a torus (the shape of the surface of a doughnut of the kind that has a hole in it), since they are fundamentally different shapes—one has a “hole” and the other does not. However, it is not easy to turn this intuition into a rigorous argument. For more on this kind of problem, see INVARIANTS [I.4 §2.2] and DIFFERENTIAL TOPOLOGY [IV.9].

6.5 Spherical Geometry

We have been steadily relaxing our requirements for two shapes to be equivalent, by allowing more and more transformations. Now let us tighten up again and look at *spherical geometry*. Here the universe is no longer \mathbb{R}^n but the n -dimensional sphere S_n , which is defined to be the surface of the $(n + 1)$ -dimensional ball, or, to put it more algebraically, the set of all points $(x_1, x_2, \dots, x_{n+1})$ in \mathbb{R}^{n+1} such that $x_1^2 + x_2^2 + \dots + x_{n+1}^2 = 1$. Just as the surface of a three-dimensional ball is two dimensional, so this set is n dimensional. We shall discuss the case $n = 2$ here, but it is easy to generalize the discussion to larger n .

The appropriate group of transformations is $SO(3)$: the group of all rotations about some axis that goes

through the origin. (One could allow reflections as well and take $O(3)$.) These are symmetries of the sphere S_2 , and that is how we regard them in spherical geometry, rather than as transformations of the whole of \mathbb{R}^3 .

Among the concepts that make sense in spherical geometry are line, distance, and angle. It may seem odd to talk about a line if one is confined to the surface of a ball, but a “spherical line” is not a line in the usual sense. Rather, it is a subset of S_2 obtained by intersecting S_2 with a plane through the origin. This produces a *great circle*, that is, a circle of radius 1, which is as large as it can be given that it lives inside a sphere of radius 1.

The reason that a great circle deserves to be thought of as some sort of line is that the shortest path between any two points x and y in S_2 will always be along a great circle, *provided that the path is confined to S_2* . This is a very natural restriction to make, since we are regarding S_2 as our “universe.” It is also a restriction of some practical relevance, since the shortest sensible route between two distant points on Earth’s surface will not be the straight-line route that burrows hundreds of miles underground.

The *distance* between two points x and y is defined to be the length of the shortest path from x to y that lies entirely in S_2 . (If x and y are opposite each other, then there are infinitely many shortest paths, all of length π , so the distance between x and y is π .) How about the *angle* between two spherical lines? Well, the lines are intersections of S_2 with two planes, so one can define it to be the angle between these two planes in the Euclidean sense. A more aesthetically pleasing way to view this, because it does not involve ideas external to the sphere, is to notice that if you look at a very small region about one of the two points where two spherical lines cross, then that portion of the sphere will be almost flat, and the lines almost straight. So you can define the angle to be the usual angle between the “limiting” straight lines inside the “limiting” plane.

Spherical geometry differs from Euclidean geometry in several interesting ways. For example, the angles of a spherical triangle always add up to *more* than 180° . Indeed, if you take as the vertices the North Pole, a point on the equator, and a second point a quarter of the way around the equator from the first, then you obtain a triangle with three right angles. The smaller a triangle, the flatter it becomes, and so the closer the sum of its angles comes to 180° . There is a beautiful theorem that gives a precise expression to this: if we switch to radians, and if we have a spherical triangle with angles α , β , and γ , then its area is $\alpha + \beta + \gamma - \pi$. (For example, this formula tells us that the triangle with three angles of $\frac{1}{2}\pi$ has area

$\frac{1}{2}\pi$, which indeed it does as the surface area of a ball of radius 1 is 4π and this triangle occupies one-eighth of the surface.)

6.6 Hyperbolic Geometry

So far, the idea of defining geometries with reference to sets of transformations may look like nothing more than a useful way to view the subject, a unified approach to what would otherwise be rather different-looking aspects. However, when it comes to hyperbolic geometry, the transformational approach becomes indispensable, for reasons that will be explained in a moment.

The group of transformations that produces hyperbolic geometry is called $\text{PSL}(2, \mathbb{R})$, the *projective special linear group* in two dimensions. One way to present this group is as follows. The *special linear group* $\text{SL}(2, \mathbb{R})$ is the set of all matrices $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with DETERMINANT [III.15] $ad - bc$ equal to 1. (These form a group because the product of two matrices with determinant 1 again has determinant 1.) To make this “projective,” one then regards each matrix A as *equivalent* to $-A$: for example, the matrices $\begin{pmatrix} 3 & -1 \\ -5 & 2 \end{pmatrix}$ and $\begin{pmatrix} -3 & 1 \\ 5 & -2 \end{pmatrix}$ are equivalent.

To get from this group to the geometry one must first interpret it as a group of transformations of some two-dimensional set of points. Once we have done this, we have what is called a *model* of two-dimensional hyperbolic geometry. The subtlety is that, unlike with spherical geometry, where the sphere was the “obvious” model, there is no single model of hyperbolic geometry that is clearly the best. (In fact, there are alternative models of spherical geometry. For example, there is a natural way of associating with each rotation of \mathbb{R}^3 a transformation of \mathbb{R}^2 with a “point at infinity” added, so the extended plane can be used as a model of spherical geometry.) The three most commonly used models of hyperbolic geometry are called the half-plane model, the disk model, and the hyperboloid model.

The *half-plane model* is the one most directly associated with the group $\text{PSL}(2, \mathbb{R})$. The set in question is the upper half-plane of the complex numbers \mathbb{C} , that is, the set of all complex numbers $z = x + yi$ such that $y > 0$. Given a matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, the corresponding transformation is the one that takes the point z to the point $(az + b)/(cz + d)$. (Notice that if we replace a, b, c , and d by their negatives, then we get the same transformation.) The condition $ad - bc = 1$ can be used to show that the transformed point will still lie in the upper half-plane, and also that the transformation can be inverted.

What this does not yet do is tell us anything about *distances*, and it is here that we need the group to “gen-

erate” the geometry. If we are to have a notion of distance d that is sensible from the perspective of our group of transformations, then it is important that the transformations should preserve it. That is, if T is one of the transformations and z and w are two points in the upper half-plane, then $d(T(z), T(w))$ should always be the same as $d(z, w)$. It turns out that there is essentially only *one* definition of distance that has this property, and that is the sense in which the group defines the geometry. (One could of course multiply all distances by some constant factor such as 3, but this would be like measuring distances in feet instead of yards, rather than a genuine difference in the geometry.)

This distance has some properties that at first seem odd. For example, a typical *hyperbolic line* takes the form of a semicircular arc with endpoints on the real axis. However, it is semicircular only from the point of view of the Euclidean geometry of \mathbb{C} : from a hyperbolic perspective it would be just as odd to regard a Euclidean straight line as straight. The reason for the discrepancy is that hyperbolic distances become larger and larger, relative to Euclidean ones, the closer you get to the real axis. To get from a point z to another point w , it is therefore shorter to take a “detour” away from the real axis, and the best detour turns out to be along an arc of the circle that goes through z and w and cuts the real axis at right angles. (If z and w are on the same vertical line, then one obtains a “degenerate circle,” namely that vertical line.) These facts are no more paradoxical than the fact that a flat map of the world involves distortions of spherical geometry, making Greenland very large, for example. The half-plane model is like a “map” of a geometric structure, the hyperbolic plane, that in reality has a very different shape.

One of the most famous properties of two-dimensional hyperbolic geometry is that it provides a geometry in which Euclid’s *parallel postulate* fails to hold. That is, it is possible to have a hyperbolic line L , a point x not on the line, and two different hyperbolic lines through x , neither of which meets L . All the other axioms of Euclidean geometry are, when suitably interpreted, true of hyperbolic geometry as well. It follows that the parallel postulate cannot be deduced from those axioms. This discovery, associated with GAUSS [VI.25], BOLYAI [VI.33], and LOBACHEVSKII [VI.30], solved a problem that had bothered mathematicians for over two thousand years.

Another property complements the result about the sum of the angles of spherical and Euclidean triangles. There is a natural notion of hyperbolic area, and the area of a hyperbolic triangle with angles α, β , and γ is $\pi - \alpha - \beta - \gamma$. Thus, in the hyperbolic plane $\alpha + \beta + \gamma$ is always

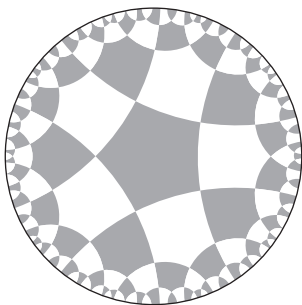


Figure 2 A tessellation of the hyperbolic disk.

less than π , and it almost equals π when the triangle is very small. These properties of angle sums reflect the fact that the sphere has positive CURVATURE [III.13], the Euclidean plane is “flat,” and the hyperbolic plane has negative curvature.

The *disk model*, conceived in a famous moment of inspiration by POINCARÉ [VI.60] as he was getting into a bus, takes as its set of points the *open unit disk* in \mathbb{C} , that is, the set D of all complex numbers with modulus less than 1. This time, a typical transformation takes the following form. One takes a real number θ and a complex number a from inside D , and sends each z in D to the point $e^{i\theta}(z - a)/(1 - \bar{a}z)$. It is not completely obvious that these transformations form a group, and still less that the group is isomorphic to $\text{PSL}(2, \mathbb{R})$. However, it turns out that the function that takes z to $-(iz + 1)/(z + i)$ maps the unit disk to the upper half-plane and vice versa. This shows that the two models give the same geometry and can be used to transfer results from one to the other.

As with the half-plane model, distances become larger, relative to Euclidean distances, as you approach the boundary of the disk: from a hyperbolic perspective, the diameter of the disk is infinite and it does not really have a boundary. Figure 2 shows a tessellation of the disk by shapes that are congruent in the sense that any one can be turned into any other by means of a transformation from the group. Thus, even though they do not look identical, within hyperbolic geometry they all have the same size and shape. Straight lines in the disk model are either arcs of (Euclidean) circles that meet the unit circle at right angles, or segments of (Euclidean) straight lines that pass through the center of the disk.

The *hyperboloid model* is the model that explains why the geometry is called hyperbolic. This time the set is the hyperboloid consisting of all points $(x, y, z) \in \mathbb{R}^3$ such that $z > 0$ and $x^2 + y^2 = 1 + z^2$. This is the hyperboloid of revolution about the z -axis of the hyperbola

$x^2 = 1 + z^2$ in the plane $y = 0$. A general transformation in the group is a sort of “rotation” of the hyperboloid, and can be built up from genuine rotations about the z -axis, and “hyperbolic rotations” of the xz -plane, which have matrices of the form

$$\begin{pmatrix} \cosh \theta & \sinh \theta \\ \sinh \theta & \cosh \theta \end{pmatrix}.$$

Just as an ordinary rotation preserves the unit circle, one of these hyperbolic rotations preserves the hyperbola $x^2 = 1 + z^2$, moving points around inside it. Again, it is not quite obvious that this gives the same group of transformations, but it does, and the hyperboloid model is equivalent to the other two.

6.7 Projective Geometry

Projective geometry is regarded by many as an old-fashioned subject, and it is no longer taught in schools, but it still has an important role to play in modern mathematics. We shall concentrate here on the *real projective plane*, but projective geometry is possible in any number of dimensions and with scalars in any field. This makes it particularly useful to algebraic geometers.

Here are two ways of regarding the projective plane. The first is that the set of points is the ordinary plane, together with a “point at infinity.” The group of transformations consists of functions known as *projections*. To understand what a projection is, imagine two planes P and P' in space, and a point x that is not in either of them. We can “project” P onto P' as follows. If a is a point in P , then its image $\phi(a)$ is the point where the line joining x to a meets P' . (If this line is parallel to P' , then $\phi(a)$ is the point at infinity of P' .) Thus, if you are at x and a picture is drawn on the plane P , then its image under the projection ϕ will be the picture drawn on P' that to you looks exactly the same. In fact, however, it will have been distorted, so the transformation ϕ has made a difference to the shape. To turn ϕ into a transformation of P itself, one can follow it by a rigid transformation that moves P' back to where P is.

Such projections do not preserve distances, but among the interesting concepts that they do preserve are points, lines, quantities known as *cross-ratios*, and, most famously, *conic sections*. A conic section is the intersection of a plane with a cone, and it can be a circle, an ellipse, a parabola, or a hyperbola. From the point of view of projective geometry, these are all the same kind of object (just as, in affine geometry, one can talk about ellipses but there is no special ellipse called a circle).

A second view of the projective plane is that it is the set of all lines in \mathbb{R}^3 that go through the origin. Since a

line is determined by the two points where it intersects the unit sphere, one can regard this set as a sphere, but with the significant difference that *opposite points are regarded as the same*—because they correspond to the same line. (This is quite hard to imagine, but not impossible. Suppose that, whatever happened on one side of the world, an identical copy of that event happened at the exactly corresponding place on the opposite side. If one was used to this situation and traveled from Paris, say, to the copy of Paris on the other side of the world, would one actually think that it was a different place? It would look the same and appear to have all the same people, and just as you arrived an identical copy of you, whom you could never meet, would be arriving in the “real” Paris. It might under such circumstances be more natural to say that there was only one Paris and only one you and that the world was not a sphere but a projective plane.)

Under this view, a typical transformation of the projective plane is obtained as follows. Take any invertible linear map, and apply it to \mathbb{R}^3 . This takes lines through the origin to lines through the origin, and can therefore be thought of as a function from the projective plane to itself. If one invertible linear map is a multiple of another, then they will have the same effect on all lines, so the resulting group of transformations is like $\text{GL}_3(\mathbb{R})$, except that all nonzero multiples of any given matrix are regarded as equivalent. This group is called the *projective special linear group* $\text{PSL}(3, \mathbb{R})$, and it is the three-dimensional equivalent of $\text{PSL}(2, \mathbb{R})$, which we have already met. Since $\text{PSL}(3, \mathbb{R})$ is bigger than $\text{PSL}(2, \mathbb{R})$, the projective plane comes with a richer set of transformations than the hyperbolic plane, which is why fewer geometrical properties are preserved. (For example, as we have seen, there is a useful notion of hyperbolic distance, but no obvious notion of projective distance.)

6.8 Lorentz Geometry

This is a geometry used in the theory of special relativity to model four-dimensional *spacetime*, otherwise known as *Minkowski space*. The main difference between it and four-dimensional Euclidean geometry is that, instead of the usual notion of distance between two points (t, x, y, z) and (t', x', y', z') , one considers the quantity

$$-(t - t')^2 + (x - x')^2 + (y - y')^2 + (z - z')^2,$$

which would be the square of the Euclidean distance were it not for the all-important minus sign before $(t - t')^2$. This reflects the fact that space and time are significantly different (though intertwined).

A *Lorentz transformation* is a linear map from \mathbb{R}^4 to \mathbb{R}^4 that preserves these “generalized distances.” Letting g be the linear map that sends (t, x, y, z) to $(-t, x, y, z)$ and letting G be the corresponding matrix (which has $-1, 1, 1, 1$ down the diagonal and 0 everywhere else), we can define a Lorentz transformation abstractly as one whose matrix Λ satisfies $\Lambda G \Lambda^T = I$, where I is the 4×4 identity matrix and Λ^T is the transpose of Λ . (The *transpose* of a matrix A is the matrix B defined by $B_{ij} = A_{ji}$.)

A point (t, x, y, z) is said to be *spacelike* if $-t^2 + x^2 + y^2 + z^2 > 0$, and *timelike* if $-t^2 + x^2 + y^2 + z^2 < 0$. If $-t^2 + x^2 + y^2 + z^2 = 0$, then the point lies in the *light cone*. All these are genuine concepts of Lorentz geometry because they are preserved by Lorentz transformations.

Lorentzian geometry is also of fundamental importance to *general relativity*, which can be thought of as the study of *Lorentzian manifolds*. These are closely related to Riemannian manifolds, which are discussed in section 6.10. For a discussion of general relativity, see GENERAL RELATIVITY AND THE EINSTEIN EQUATIONS [IV.17].

6.9 Manifolds and Differential Geometry

To somebody who has not been taught otherwise, it is natural to think that Earth is flat, or rather that it consists of a flat surface on top of which there are buildings, mountains, and so on. However, we now know that it is in fact more like a sphere, appearing to be flat only because it is so large. There are various kinds of evidence for this. One is that if you stand on a cliff by the sea then you can see a definite horizon, not too far away, over which ships disappear. This would be hard to explain if Earth were genuinely flat. Another is that if you travel far enough in what feels like a straight line then you eventually get back to where you started. A third is that if you travel along a triangular route and the triangle is a large one, then you will be able to detect that its three angles add up to more than 180° .

It is also very natural to believe that the geometry that best models that of the universe is three-dimensional Euclidean geometry, or what one might think of as “normal” geometry. However, this could be just as much of a mistake as believing that two-dimensional Euclidean geometry is the best model for Earth’s surface.

Indeed, one can immediately improve on it by considering Lorentz geometry as a model of spacetime, but even if there were no theory of special relativity, our astronomical observations would give us no particular reason to suppose that Euclidean geometry was the best

model for the universe. Why should we be so sure that we would not obtain a better model by taking the three-dimensional surface of a very large four-dimensional sphere? This might feel like “normal” space in just the way that the surface of Earth feels like a “normal” plane unless you travel large distances. Perhaps if you traveled far enough in a rocket without changing your course then you would end up where you started.

It is easy to describe “normal” space mathematically: one just associates with each point in space a triple of coordinates (x, y, z) in the usual way. How might we describe a huge “spherical” space? It is slightly harder, but not much: one can give each point *four* coordinates (x, y, z, w) but add the condition that these must satisfy the equation $x^2 + y^2 + z^2 + w^2 = R^2$ for some fixed R that we think of as the “radius” of the universe. This describes the three-dimensional surface of a four-dimensional sphere of radius R in just the same way that the equation $x^2 + y^2 + z^2 = R^2$ describes the two-dimensional surface of a three-dimensional sphere of radius R .

A possible objection to this approach is that it seems to rely on the rather implausible idea that the universe lives in some larger unobserved four-dimensional space. However, this objection can be answered. The object we have just defined, the 3-sphere S_3 , can also be described in what is known as an *intrinsic* way: that is, without reference to some surrounding space. The easiest way to see this is to discuss the 2-sphere first, in order to draw an analogy.

Let us therefore imagine a planet covered with calm water. If you drop a large rock into the water at the North Pole, a wave will propagate out in a circle of ever-increasing radius. (At any one moment, it will be a circle of constant latitude.) In due course, however, this circle will reach the equator, after which it will start to *shrink*, until eventually the whole wave reaches the South Pole at once, in a sudden burst of energy.

Now imagine setting off a three-dimensional wave in space—it could, for example, be a light wave caused by the switching on of a bright light. The front of this wave would now be not a circle but an ever-expanding spherical surface. It is logically possible that this surface could expand until it became very large and then contract again, not by shrinking back to where it started, but by turning itself inside out, so to speak, and shrinking to another point on the opposite side of the universe. (Notice that in the two-dimensional example, what you want to call the inside of the circle changes when the circle passes the equator.) With a bit of effort, one can visualize this possibility, and there is no need

to appeal to the existence of a fourth dimension in order to do so. More to the point, this account can be turned into a mathematically coherent and genuinely three-dimensional description of the 3-sphere.

A different and more general approach is to use what is called an *atlas*. An atlas of the world (in the normal, everyday sense) consists of a number of flat pages, together with an indication of their *overlaps*: that is, of how parts of some pages correspond to parts of others. Now, although such an atlas is mapping out an external object that lives in a three-dimensional universe, the spherical geometry of Earth’s surface can be read off from the atlas alone. It may be much less convenient to do this but it is possible: rotations, for example, might be described by saying that such-and-such a part of page 17 moved to a similar but slightly distorted part of page 24, and so on.

Not only is this possible, but one can *define* a surface by means of two-dimensional atlases. For example, there is a mathematically neat “atlas” of the 2-sphere that consists of just two pages, both of them circular. One is a map of the Northern Hemisphere plus a little bit of the Southern Hemisphere near the equator (to provide a small overlap) and the other is a map of the Southern Hemisphere with a bit of the Northern Hemisphere. Because these maps are flat, they necessarily involve some distortion, but one can specify what this distortion is.

The idea of an atlas can easily be generalized to three dimensions. A “page” now becomes a portion of three-dimensional space. The technical term is not “page” but “chart,” and a three-dimensional atlas is a collection of charts, again with specifications of which parts of one chart correspond to which parts of another. A possible atlas of the 3-sphere, generalizing the simple atlas of the 2-sphere just discussed, consists of two solid three-dimensional balls. There is a correspondence between points toward the edge of one of these balls and points toward the edge of the other, and this can be used to describe the geometry: as you travel toward the edge of one ball you find yourself in the overlapping region, so you are also in the other ball. As you go further, you are off the map as far as the first ball is concerned, but the second ball has by that stage taken over.

The 2-sphere and the 3-sphere are basic examples of *manifolds*. Other examples that we have already met in this section are the torus and the projective plane. Informally, a d -dimensional manifold, or d -manifold, is any geometrical object M with the property that every point x in M is surrounded by what feels like a portion of d -dimensional Euclidean space. So, because small parts of

a sphere, torus, or projective plane are very close to planar, they are all 2-manifolds, though when the dimension is two the word *surface* is more usual. (However, it is important to remember that a “surface” need not be the surface *of* anything.) Similarly, the 3-sphere is a 3-manifold.

The formal definition of a manifold uses the idea of atlases: indeed, one says that the atlas *is* a manifold. This is a typical mathematician’s use of the word “is,” and it should not be confused with the normal use. In practice, it is unusual to think of a manifold as a collection of charts with rules for how parts of them correspond, but the definition in terms of charts and atlases turns out to be the most convenient when one wishes to reason about manifolds in general rather than discussing specific examples. For the purposes of this book, it may be better to think of a d -manifold in the “extrinsic” way that we first thought about the 3-sphere: as a d -dimensional “hypersurface” living in some higher-dimensional space. Indeed, there is a famous theorem of Nash that states that all manifolds arise in this way. Note, however, that it is not always easy to find a simple formula for defining such a hypersurface. For example, while the 2-sphere is described by the simple formula $x^2 + y^2 + z^2 = 1$ and the torus by the slightly more complicated and more artificial formula $(r - 2)^2 + z^2 = 1$, where r is shorthand for $\sqrt{x^2 + y^2}$, it is not easy to come up with a formula that describes a two-holed torus. Even the usual torus is far more easily described using quotients, as we did in section 3.3. Quotients can also be used to define a two-holed torus (see FUCHSIAN GROUPS [III.28]), and the reason one is confident that the result is a manifold is that every point has a small neighborhood that looks like a small part of the Euclidean plane. In general, a d -dimensional manifold can be thought of as any construction that gives rise to an object that is “locally like Euclidean space of d dimensions.”

An extremely important feature of manifolds is that calculus is possible for functions defined on them. Roughly speaking, if M is a manifold and f is a function from M to \mathbb{R} , then to see whether f is differentiable at a point x in M you first find a chart that contains x (or a representation of it), and regard f as a function defined on the chart instead. Since the chart is a portion of the d -dimensional Euclidean space \mathbb{R}^d and we can differentiate functions defined on such sets, the notion of differentiability now makes sense for f . Of course, for this definition to work for the manifold, it is important that if x belongs to two overlapping charts, then the answer will be the same for both. This is guaranteed if the function that gives the correspondence between the overlap-

ping parts (known as a *transition function*) is itself differentiable. Manifolds with this property are called *differentiable manifolds*: manifolds for which the transition functions are continuous but not necessarily differentiable are called *topological manifolds*. The availability of calculus makes the theory of differentiable manifolds very different from that of topological manifolds.

The above ideas generalize easily from real-valued functions to functions from M to \mathbb{R}^d , or from M to M' , where M' is another manifold. However, it is easier to judge whether a function defined on a manifold is differentiable than it is to say what the derivative is. The derivative at some point x of a function from \mathbb{R}^n to \mathbb{R}^m is a linear map, and so is the derivative of a function defined on a manifold. However, the domain of the linear map is not the manifold itself, which is not usually a vector space, but rather the so-called *tangent space* at the point x in question.

For more details on this and on manifolds in general, see DIFFERENTIAL TOPOLOGY [IV.9].

6.10 Riemannian Metrics

Suppose you are given two points P and Q on a sphere. How do you determine the distance between them? The answer depends on how the sphere is defined. If it is the set of all points (x, y, z) such that $x^2 + y^2 + z^2 = 1$ then P and Q are points in \mathbb{R}^3 . One can therefore use the Pythagorean theorem to calculate the distance between them. For example, the distance between the points $(1, 0, 0)$ and $(0, 1, 0)$ is $\sqrt{2}$.

However, do we really want to measure the length of the line segment PQ ? This segment does not lie in the sphere itself, so to use it as a means of defining length does not sit at all well with the idea of a manifold as an intrinsically defined object. Fortunately, as we saw earlier in the discussion of spherical geometry, there is another natural definition that avoids this problem: we can define the distance between P and Q as the length of the shortest path from P to Q that lies entirely within the sphere.

Now let us suppose that we wish to talk more generally about distances between points in manifolds. If the manifold is presented to us as a hypersurface in some bigger space, then we can use lengths of shortest paths as we did in the sphere. But suppose that the manifold is presented differently and all we have is a way of demonstrating that every point is contained in a chart—that is, has a neighborhood that can be associated with a portion of d -dimensional Euclidean space. (For the purposes of this discussion, nothing is lost if one takes d to be

2 throughout, in which case there is a correspondence between the neighborhood and a portion of the plane.) One idea is to define the distance between the two points to be the distance between the corresponding points in the chart, but this raises at least three problems.

The first is that the points P and Q that we are looking at might belong to different charts. This, however, is not too much of a problem, since all we actually need to do is calculate lengths of paths, and that can be done provided we have a way of defining distances between points that are very close together, in which case we can find a single chart that contains them both.

The second problem, which is much more serious, is that for any one manifold there are many ways of choosing the charts, so this idea does not lead to a single notion of distance for the manifold. Worse still, even if one fixes one set of charts, these charts will overlap, and it may not be possible to make the notions of distance compatible where the overlap occurs.

The third problem is related to the second. The surface of a sphere is curved, whereas the charts of any atlas (in either the everyday or the mathematical sense) are flat. Therefore, the distances in the charts cannot correspond exactly to the lengths of shortest paths in the sphere itself.

The single most important moral to draw from the above problems is that if we wish to define a notion of distance for a given manifold, we have a great deal of choice about how to do so. Very roughly, a Riemannian metric is a way of making such a choice.

A little less roughly, a *metric* means a sensible notion of distance (the precise definition can be found in [III.58]). A Riemannian metric is a way of determining infinitesimal distances. These infinitesimal distances can be used to calculate lengths of paths, and then the distance between two points can be defined as the length of the shortest path between them. To see how this is done, let us first think about lengths of paths in the ordinary Euclidean plane. Suppose that (x, y) belongs to a path and $(x + \delta x, y + \delta y)$ is another point on the path, very close to (x, y) . Then the distance between the two points is $\sqrt{\delta x^2 + \delta y^2}$. To calculate the length of a sufficiently smooth path, one can choose a large number of points along the path, each one very close to the next, and add up their distances. This gives a good approximation, and one can make it better and better by taking more and more points.

In practice, it is easier to work out the length using calculus. A path itself can be thought of as a moving point $(x(t), y(t))$ that starts when $t = 0$ and ends when $t = 1$. If δt is very small, then $x(t + \delta t)$ is approximately $x(t) +$

$x'(t)\delta t$ and $y(t + \delta t)$ is approximately $y(t) + y'(t)\delta t$. Therefore, the distance between $(x(t), y(t))$ and $(x(t + \delta t), y(t + \delta t))$ is approximately $\delta t\sqrt{x'(t)^2 + y'(t)^2}$, by the Pythagorean theorem. Therefore, letting δt go to zero and integrating all the infinitesimal distances along the path, we obtain the formula

$$\int_0^1 \sqrt{x'(t)^2 + y'(t)^2} dt$$

for the length of the path. Notice that if we write $x'(t)$ and $y'(t)$ as dx/dt and dy/dt , then we can rewrite $\sqrt{x'(t)^2 + y'(t)^2} dt$ as $\sqrt{dx^2 + dy^2}$, which is the infinitesimal version of our earlier expression $\sqrt{\delta x^2 + \delta y^2}$. We have just defined a Riemannian metric, which is usually denoted by $dx^2 + dy^2$. This can be thought of as the square of the distance between (x, y) and the infinitesimally close point $(x + dx, y + dy)$.

If we want to, we can now prove that the shortest path between two points (x_0, y_0) and (x_1, y_1) is a straight line, which will tell us that the distance between them is $\sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}$. (A proof can be found in VARIATIONAL METHODS [III.94].) However, since we could have just used this formula to begin with, this example does not really illustrate what is distinctive about Riemannian metrics. To do that, let us give a more precise definition of the disk model for hyperbolic geometry, which was discussed in section 6.6. There it was stated that distances become larger, relative to Euclidean distances, as one approaches the edge of the disk. A more precise definition is that the *open unit disk* is the set of all points (x, y) such that $x^2 + y^2 < 1$ and that the Riemannian metric on this disk is given by the expression $(dx^2 + dy^2)/(1 - x^2 - y^2)$. This is how we *define* the square of the distance between (x, y) and $(x + dx, y + dy)$. Equivalently, the length of a path $(x(t), y(t))$ with respect to this Riemannian metric is defined as

$$\int_0^1 \sqrt{\frac{x'(t)^2 + y'(t)^2}{1 - x(t)^2 - y(t)^2}} dt.$$

More generally, a *Riemannian metric* on a portion of the plane is an expression of the form

$$E(x, y) dx^2 + 2F(x, y) dx dy + G(x, y) dy^2$$

that is used to calculate infinitesimal distances and hence lengths of paths. (In the disk model we took $E(x, y)$ and $G(x, y)$ to be $1/(1 - x^2 - y^2)$ and $F(x, y)$ to be 0.) It is important for these distances to be positive, which will turn out to be the case provided that $E(x, y)G(x, y) - F(x, y)^2$ is always positive. One also needs the functions E , F , and G to satisfy certain smoothness conditions.

This definition generalizes straightforwardly to more dimensions. In n dimensions we must specify the squared distance between (x_1, \dots, x_n) and $(x_1 + dx_1, \dots, x_n + dx_n)$, using an expression of the form

$$\sum_{i,j=1}^n F_{ij}(x_1, \dots, x_n) dx_i dx_j.$$

The numbers $F_{ij}(x_1, \dots, x_n)$ form an $n \times n$ matrix that depends on the point (x_1, \dots, x_n) . This matrix is required to be symmetric and positive definite, which means that $F_{ij}(x_1, \dots, x_n)$ should always equal $F_{ji}(x_1, \dots, x_n)$ and the expression that determines the squared distance should always be positive. It should also depend smoothly on the point (x_1, \dots, x_n) .

Finally, now that we know how to define many different Riemannian metrics on portions of Euclidean space, we have many potential ways to define metrics on the charts that we use to define a manifold. A Riemannian metric on a *manifold* is a way of choosing compatible Riemannian metrics on the charts, where “compatible” means that wherever two charts overlap the distances should be the same. As mentioned earlier, once one has done this, one can define the distance between two points to be the length of a shortest path between them.

Given a Riemannian metric on a manifold, it is possible to define many other concepts, such as angles and volumes. It is also possible to define the important concept of *curvature*, which is discussed in RICCI FLOW [III.80]. Another important definition is that of a *geodesic*, which is the analogue for Riemannian geometry of a straight line in Euclidean geometry. A curve C is a geodesic if, given any two points P and Q on C that are sufficiently close, the shortest path from P to Q is part of C . For example, the geodesics on the sphere are the great circles.

As should be clear by now from the above discussion, on any given manifold there is a multitude of possible Riemannian metrics. A major theme in Riemannian geometry is to choose one that is “best” in some way. For example, on the sphere, if we take the obvious definition of the length of a path, then the resulting metric is particularly symmetric, and this is a highly desirable property. In particular, with this Riemannian metric the curvature of the sphere is the same everywhere. More generally, one searches for extra conditions to impose on Riemannian metrics. Ideally, these conditions should be strong enough that there is just one Riemannian metric that satisfies them, or at least that the family of such metrics should be very small.

I.4 The General Goals of Mathematical Research

The previous article introduced many concepts that appear throughout mathematics. This one discusses what mathematicians do with those concepts, and the sorts of questions they ask about them.

1 Solving Equations

As we have seen in earlier articles, mathematics is full of objects and structures (of a mathematical kind), but they do not simply sit there for our contemplation: we also like to *do* things to them. For example, given a number, there will be contexts in which we want to double it, or square it, or work out its reciprocal; given a suitable function, we may wish to differentiate it; given a geometrical shape, we may wish to transform it; and so on.

Transformations like these give rise to a never-ending source of interesting problems. If we have defined some mathematical process, then a rather obvious mathematical project is to invent techniques for carrying it out. This leads to what one might call *direct* questions about the process. However, there is also a deeper set of *inverse* questions, which take the following form. Suppose you are told what process has been carried out and what answer it has produced. Can you then work out what the mathematical object was that the process was applied to? For example, suppose I tell you that I have just taken a number and squared it, and that the result was 9. Can you tell me the original number?

In this case the answer is more or less yes: it must have been 3, except that if negative numbers are allowed, then another solution is -3 .

If we want to talk more formally, then we say that we have been examining the equation $x^2 = 9$, and have discovered that there are two solutions. This example raises three issues that appear again and again.

- Does a given equation have any solutions?
- If so, does it have exactly one solution?
- What is the set in which solutions are required to live?

The first two concerns are known as the *existence* and the *uniqueness* of solutions. The third does not seem particularly interesting in the case of the equation $x^2 = 9$, but in more complicated cases, such as partial differential equations, it can be a subtle and important question.

To use more abstract language, suppose that f is a FUNCTION [I.2 §2.2] and we are faced with a statement of the form $f(x) = y$. The direct question is to work out y given what x is. The inverse question is to work out x given what y is: this would be called solving the equation $f(x) = y$. Not surprisingly, questions about the solutions of an equation of this form are closely related to questions about the invertibility of the function f , which were discussed in [I.2]. Because x and y can be very much more general objects than numbers, the notion of solving equations is itself very general, and for that reason it is central to mathematics.

1.1 Linear Equations

The very first equations a schoolchild meets will typically be ones like $2x + 3 = 17$. To solve simple equations like this, one treats x as an unknown number that obeys the usual rules of arithmetic. By exploiting these rules one can transform the equation into something much simpler: subtracting 3 from both sides we learn that $2x = 14$, and dividing both sides of this new equation by 2 we then discover that $x = 7$. If we are very careful, we will notice that all we have shown is that *if* there is some number x such that $2x + 3 = 17$ *then* x must be 7. What we have not shown is that there is any such x . So strictly speaking there is a further step of checking that $2 \times 7 + 3 = 17$. This will obviously be true here, but the corresponding assertion is not always true for more complicated equations so this final step can be important.

The equation $2x + 3 = 17$ is called “linear” because the function f we have performed on x (to multiply it by 2 and add 3) is a linear one, in the sense that its graph is a straight line. As we have just seen, linear equations involving a single unknown x are easy to solve, but matters become considerably more sophisticated when one starts to deal with more than one unknown. Let us look at a typical example of an equation in two unknowns, the equation $3x + 2y = 14$. This equation has many solutions: for any choice of y you can set $x = (14 - 2y)/3$ and you have a pair (x, y) that satisfies the equation. To make it harder, one can take a second equation as well, $5x + 3y = 22$, say, and try to solve the two equations *simultaneously*. Then, it turns out, there is just one solution, namely $x = 2$ and $y = 4$. Typically, two linear equations in two unknowns have exactly one solution, just as these two do, which is easy to see if one thinks about the situation geometrically. An equation of the form $ax + by = c$ is the equation of a straight line in the xy -plane. Two lines normally meet in a single point,

the exceptions being when they are identical, in which case they meet in infinitely many points, or parallel but not identical, in which case they do not meet at all.

If one has several equations in several unknowns, it can be conceptually simpler to think of them as one equation in one unknown. This sounds impossible, but it is perfectly possible if the new unknown is allowed to be a more complicated object. For example, the two equations $3x + 2y = 14$ and $5x + 3y = 22$ can be rewritten as the following single equation involving matrices and vectors:

$$\begin{pmatrix} 3 & 2 \\ 5 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 14 \\ 22 \end{pmatrix}.$$

If we let A stand for the matrix, \mathbf{x} for the unknown column vector, and \mathbf{b} for the known one, then this equation becomes simply $A\mathbf{x} = \mathbf{b}$, which looks much less complicated, even if in fact all we have done is hidden the complication behind our notation.

There is more to this process, however, than sweeping dirt under the carpet. While the simpler notation conceals many of the specific details of the problem, it also *reveals* very clearly what would otherwise be obscured: that we have a linear map from \mathbb{R}^2 to \mathbb{R}^2 and we want to know which vectors \mathbf{x} , if any, map to the vector \mathbf{b} . When faced with a particular set of simultaneous equations, this reformulation does not make much difference—the calculations we have to do are the same—but when we wish to reason more generally, either directly about simultaneous equations or about other problems where they arise, it is much easier to think about a matrix equation with a single unknown vector than about a collection of simultaneous equations in several unknown numbers. This phenomenon occurs throughout mathematics and is a major reason for the study of high-dimensional spaces.

1.2 Polynomial Equations

We have just discussed the generalization of linear equations from one variable to several variables. Another direction in which one can generalize them is to think of linear functions as polynomials of degree 1 and consider functions of higher degree. At school, for example, one learns how to solve *quadratic* equations, such as $x^2 - 7x + 12 = 0$. More generally, a *polynomial equation* is one of the form

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0 = 0.$$

To solve such an equation means to find a value of x for which the equation is true (or, better still, all such values). This may seem an obvious thing to say until one considers a very simple example such as the equation

$x^2 - 2 = 0$, or equivalently $x^2 = 2$. The solution to this is, of course, $x = \pm\sqrt{2}$. What, though, is $\sqrt{2}$? It is defined to be the positive number that squares to 2, but it does not seem to be much of a “solution” to the equation $x^2 = 2$ to say that x is plus or minus the positive number that squares to 2. Neither does it seem entirely satisfactory to say that $x = 1.4142135\dots$, since this is just the beginning of a calculation that never finishes and does not result in any discernible pattern.

There are two lessons that can be drawn from this example. One is that what matters about an equation is often the *existence* and *properties* of solutions and not so much whether one can find a formula for them. Although we do not appear to learn anything when we are told that the solutions to the equation $x^2 = 2$ are $x = \pm\sqrt{2}$, this assertion does contain within it a fact that is not wholly obvious: that the number 2 has a square root. This is usually presented as a consequence of the *intermediate value theorem* (or another result of a similar nature), which states that if f is a continuous real-valued function and $f(a)$ and $f(b)$ lie on either side of 0, then somewhere between a and b there must be a c such that $f(c) = 0$. This result can be applied to the function $f(x) = x^2 - 2$, since $f(1) = -1$ and $f(2) = 2$. Therefore, there is some x between 1 and 2 such that $x^2 - 2 = 0$, that is, $x^2 = 2$. For many purposes, the mere existence of this x is enough, together with its defining properties of being positive and squaring to 2.

A similar argument tells us that all positive real numbers have positive square roots. But the picture changes when we try to solve more complicated quadratic equations. Then we have two choices. Consider, for example, the equation $x^2 - 6x + 7 = 0$. We could note that $x^2 - 6x + 7$ is -1 when $x = 4$ and 2 when $x = 5$ and deduce from the intermediate value theorem that the equation has some solution between 4 and 5. However, we do not learn as much from this as if we complete the square, rewriting $x^2 - 6x + 7$ as $(x - 3)^2 - 2$. This allows us to rewrite the equation as $(x - 3)^2 = 2$, which has the two solutions $x = 3 \pm \sqrt{2}$. We have already established that $\sqrt{2}$ exists and lies between 1 and 2, so not only do we have a solution of $x^2 - 6x + 7 = 0$ that lies between 4 and 5, but we can see that it is closely related to, indeed built out of, the solution to the equation $x^2 = 2$. This demonstrates a second important aspect of equation solving, which is that in many instances the explicit solubility of an equation is a *relative* notion. If we are given a solution to the equation $x^2 = 2$, we do not need any *new* input from the intermediate value theorem to solve the more complicated equation $x^2 - 6x + 7 = 0$: all we need is some algebra. The solution, $x = 3 \pm \sqrt{2}$, is given by an

explicit expression, but inside that expression we have $\sqrt{2}$, which is *not* defined by means of an explicit formula but as a real number, with certain properties, that we can prove to exist.

Solving polynomial equations of higher degree is markedly more difficult than solving quadratics, and raises fascinating questions. In particular, there are complicated formulas for the solutions of cubic and quartic equations, but the problem of finding corresponding formulas for quintic and higher-degree equations became one of the most famous unsolved problems in mathematics, until ABEL [VI.32] and GALOIS [VI.40] showed that it could not be done. For more details about these matters see THE INSOLUBILITY OF THE QUINTIC [V.24]. For another article related to polynomial equations see THE FUNDAMENTAL THEOREM OF ALGEBRA [V.15].

1.3 Polynomial Equations in Several Variables

Suppose that we are faced with an equation such as

$$x^3 + y^3 + z^3 = 3x^2y + 3y^2z + 6xyz.$$

We can see straight away that there will be many solutions: if you fix x and y , then the equation is a cubic polynomial in z , and all cubics have at least one (real) solution. Therefore, for every choice of x and y there is some z such that the triple (x, y, z) is a solution of the above equation.

Because the formula for the solution of a general cubic equation is rather complicated, a precise specification of the set of all triples (x, y, z) that solve the equation may not be very enlightening. However, one can learn a lot by regarding this solution set as a geometric object—a two-dimensional surface in space, to be precise—and to ask *qualitative* questions about it. One might, for instance, wish to understand roughly what shape it is. Questions of this kind can be made precise using the language and concepts of TOPOLOGY [I.3 §6.4].

One can of course generalize further and consider simultaneous solutions to several polynomial equations. Understanding the solution sets of such systems of equations is the province of ALGEBRAIC GEOMETRY [IV.7].

1.4 Diophantine Equations

As has been mentioned, the answer to the question of whether a particular equation has a solution varies according to where the solution is allowed to be. The equation $x^2 + 3 = 0$ has no solution if x is required to be real, but in the complex numbers it has the two solutions $x = \pm i\sqrt{3}$. The equation $x^2 + y^2 = 11$ has infinitely

many solutions if we are looking for x and y in the real numbers, but none if they have to be integers.

This last example is a typical *Diophantine equation*, the name given to an equation if one is looking for integer solutions. The most famous Diophantine equation is the Fermat equation $x^n + y^n = z^n$, which is now known, thanks to Andrew Wiles, to have no positive integer solutions if n is greater than 2. (See FERMAT'S LAST THEOREM [V.12]. By contrast, the equation $x^2 + y^2 = z^2$ has infinitely many solutions.) A great deal of modern ALGEBRAIC NUMBER THEORY [IV.3] is concerned with Diophantine equations, either directly or indirectly. As with equations in the real and complex numbers, it is often fruitful to study the structure of sets of solutions to Diophantine equations: this investigation belongs to the area known as ARITHMETIC GEOMETRY [IV.6].

A notable feature of Diophantine equations is that they tend to be extremely difficult. It is therefore natural to wonder whether there could be a systematic approach to them. This question was the tenth in a famous list of problems asked by HILBERT [VI.62] in 1900. It was not until 1970 that Yuri Matiyasevitch, building on work by Martin Davis, Julia Robinson, and Hilary Putnam, proved that the answer was no. (This is discussed further in THE INSOLUBILITY OF THE HALTING PROBLEM [V.23].)

An important step in the solution was taken in 1936, by Church and TURING [VI.92]. This was to make precise the notion of a “systematic approach,” by formalizing (in two different ways) the notion of an algorithm (see ALGORITHMS [II.4 §3] and COMPUTATIONAL COMPLEXITY [IV.21 §1]). It was not easy to do this in the pre-computer age, but now we can restate the solution of Hilbert's tenth problem as follows: there is no computer program that can take as its input any Diophantine equation, and without fail print “YES” if it has a solution and “NO” otherwise.

What does this tell us about Diophantine equations? We can no longer dream of a final theory that will encompass them all, so instead we are forced to restrict our attention to individual equations or special classes of equations, continually developing different methods for solving them. This would make them uninteresting after the first few, were it not for the fact that specific Diophantine equations have remarkable links with very general questions in other parts of mathematics. For example, equations of the form $y^2 = f(x)$, where $f(x)$ is a cubic polynomial in x , may look rather special, but in fact the ELLIPTIC CURVES [III.21] that they define are central to modern number theory, including the proof of Fermat's last theorem. Of course, Fermat's last theorem is itself a Diophantine equation, but its study has led to

major developments in other parts of number theory. The correct moral to draw is perhaps this: solving a particular Diophantine equation is fascinating and worthwhile if, as is often the case, the result is more than a mere addition to the list of equations that have been solved.

1.5 Differential Equations

So far, we have looked at equations where the unknown is either a number or a point in n -dimensional space (that is, a sequence of n numbers). To generate these equations, we took various combinations of the basic arithmetical operations and applied them to our unknowns.

Here, for comparison, are two well-known differential equations, the first “ordinary” and the second “partial”:

$$\begin{aligned} \frac{d^2x}{dt^2} + k^2x &= 0, \\ \frac{\partial T}{\partial t} &= \kappa \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} \right). \end{aligned}$$

The first is the equation for simple harmonic motion, which has the general solution $x(t) = A \sin kt + B \cos kt$; the second is the heat equation, which was discussed in SOME FUNDAMENTAL MATHEMATICAL DEFINITIONS [I.3 §5.4].

For many reasons, differential equations represent a jump in sophistication. One is that the unknowns are *functions*, which are much more complicated objects than numbers or n -dimensional points. (For example, the first equation above asks what function x of t has the property that if you differentiate it twice then you get $-k^2$ times the original function.) A second is that the basic operations one performs on functions include differentiation and integration, which are considerably less “basic” than addition and multiplication. A third is that differential equations that can be solved in “closed form,” that is, by means of a formula for the unknown function f , are the exception rather than the rule, even when the equations are natural and important.

Consider again the first equation above. Suppose that, given a function f , we write $\phi(f)$ for the function $(d^2f/dt^2) + k^2f$. Then ϕ is a linear map, in the sense that $\phi(f + g) = \phi(f) + \phi(g)$ and $\phi(af) = a\phi(f)$ for any constant a . This means that the differential equation can be regarded as something like a matrix equation, but generalized to infinitely many dimensions. The heat equation has the same property: if we define $\psi(T)$ to be

$$\frac{\partial T}{\partial t} - \kappa \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} \right),$$

then ψ is another linear map. Such differential equations are called *linear*, and the link with linear algebra makes them markedly easier to solve. (A very useful tool for this is THE FOURIER TRANSFORM [III.27].)

What about the more typical equations, the ones that cannot be solved in closed form? Then the focus shifts once again toward establishing whether or not solutions *exist*, and if so what *properties* they have. As with polynomial equations, this can depend on what you count as an allowable solution. Sometimes we are in the position we were in with the equation $x^2 = 2$: it is not too hard to prove that solutions exist and all that is left to do is name them. A simple example is the equation $dy/dx = e^{-x^2}$. In a certain sense, this cannot be solved: it can be shown that there is no function built out of polynomials, EXPONENTIALS [III.25], and TRIGONOMETRIC FUNCTIONS [III.93] that differentiates to e^{-x^2} . However, in another sense the equation is easy to solve—all you have to do is integrate the function e^{-x^2} . The resulting function (when divided by $\sqrt{2\pi}$) is the NORMAL DISTRIBUTION [III.73 §5] function. The normal distribution is of fundamental importance in probability, so the function is given a name, ϕ .

In most situations, there is no hope of writing down a formula for a solution, even if one allows oneself to integrate “known” functions. A famous example is the so-called THREE-BODY PROBLEM [V.36]: given three bodies moving in space and attracted to each other by gravitational forces, how will they continue to move? Using Newton’s laws, one can write down some differential equations that describe this situation. NEWTON [VI.13] solved the corresponding equations for two bodies, and thereby explained why planets move in elliptical orbits around the Sun, but for three or more bodies they proved very hard indeed to solve. It is now known that there was a good reason for this: the equations can lead to chaotic behavior (see DYNAMICS [IV.15] for more about chaos). However, this opens up a new and very interesting avenue of research into questions of chaos and stability.

Sometimes there are ways of proving that solutions exist even if they cannot be easily specified. Then one may ask not for precise formulas, but for general descriptions. For example, if the equation has a time dependence (as, for instance, the heat equation and wave equations have), one can ask whether solutions tend to decay over time, or blow up, or remain roughly the same. These more qualitative questions concern what is known as *asymptotic behavior*, and there are techniques for answering some of them even when a solution is not given by a tidy formula.

As with Diophantine equations, there are some special and important classes of partial differential equations, including nonlinear ones, that *can* be solved exactly. This gives rise to a very different style of research: again one is interested in properties of solutions, but now these properties may be more algebraic in nature, in the sense that exact formulas will play a more important role. See LINEAR AND NONLINEAR WAVES AND SOLITONS [III.51].

2 Classifying

If one is trying to understand a new mathematical structure, such as a GROUP [I.3 §2.1] or a MANIFOLD [I.3 §6.9], one of the first tasks is to come up with a good supply of examples. Sometimes examples are very easy to find, in which case there may be a bewildering array of them that cannot be put into any sort of order. Often, however, the conditions that an example must satisfy are quite stringent, and then it may be possible to come up with something like an infinite list that includes every single one. For example, it can be shown that any VECTOR SPACE [I.3 §2.3] of dimension n over a field \mathbb{F} is isomorphic to \mathbb{F}^n . This means that just one positive integer, n , is enough to determine the space completely. In this case our “list” will be $\{0\}, \mathbb{F}, \mathbb{F}^2, \mathbb{F}^3, \mathbb{F}^4, \dots$. In such a situation we say that we have a *classification* of the mathematical structure in question.

Classifications are very useful because if we can classify a mathematical structure then we have a new way of proving results about that structure: instead of deducing a result from the axioms that the structure is required to satisfy, we can simply check that it holds for every example on the list, confident in the knowledge that we have thereby proved it in general. This is not always easier than the more abstract, axiomatic approach, but it certainly is sometimes. Indeed, there are several results proved using classifications that nobody knows how to prove in any other way. More generally, the more examples you know of a mathematical structure, the easier it is to think about that structure—testing hypotheses, finding counterexamples, and so on. If you know *all* the examples of the structure, then for some purposes your understanding is complete.

2.1 Identifying Building Blocks and Families

There are two situations that typically lead to interesting classification theorems. The boundary between them is somewhat blurred, but the distinction is clear enough to be worth making, so we shall discuss them separately in this subsection and the next.

As an example of the first kind of situation, let us look at objects called *regular polytopes*. Polytopes are polygons, polyhedra, and their higher-dimensional generalizations. The regular polygons are those for which all sides have the same length and all angles are equal, and the regular polyhedra are those for which all faces are congruent regular polygons and every vertex has the same number of edges coming out of it. More generally, a higher-dimensional polytope is regular if it is as symmetrical as possible, though the precise definition of this is somewhat complicated. (Here, in three dimensions, is a definition that turns out to be equivalent to the one just given but easier to generalize. A *flag* is a triple (v, e, f) where v is a vertex of the polyhedron, e is an edge containing v , and f is a face containing e . A polyhedron is regular if for any two flags (v, e, f) and (v', e', f') there is a symmetry of the polyhedron that takes v to v' , e to e' , and f to f' .)

It is easy to see what the regular polygons are in two dimensions: for every k greater than 2 there is exactly one regular k -gon and that is all there is. In three dimensions, the regular polyhedra are the famous *Platonic solids*, that is, the tetrahedron, the cube, the octahedron, the dodecahedron, and the icosahedron. It is not too hard to see that there cannot be any more regular polyhedra, since there must be at least three faces meeting at every vertex, and the angles at that vertex must add up to less than 360° . This constraint means that the only possibilities for the faces at a vertex are three, four, or five triangles, three squares, or three pentagons. These give the tetrahedron, the octahedron, the icosahedron, the cube, and the dodecahedron, respectively.

Some of the polygons and polyhedra just defined have natural higher-dimensional analogues. For example, if you take $n + 1$ points in \mathbb{R}^n all at the same distance from one another, then they form the vertices of a *regular simplex*, which is an equilateral triangle or regular tetrahedron when $n = 2$ or 3 . The set of all points (x_1, x_2, \dots, x_n) with $0 \leq x_i \leq 1$ for every i forms the n -dimensional analogue of a unit square or cube. The octahedron can be defined as the set of all points (x, y, z) in \mathbb{R}^3 such that $|x| + |y| + |z| \leq 1$, and the analogue of this in n dimensions is the set of all points (x_1, x_2, \dots, x_n) such that $|x_1| + \dots + |x_n| \leq 1$.

It is not obvious how the dodecahedron and icosahedron would lead to infinite families of regular polytopes, and it turns out that they do not. In fact, apart from three more examples in four dimensions, the above polytopes constitute a complete list. These three examples are quite remarkable. One of them has 120 “three-dimensional faces,” each of which is a regular dodec-

ahedron. It has a so-called dual, which has 600 regular tetrahedra as its “faces.” The third example can be described in terms of coordinates: its vertices are the sixteen points of the form $(\pm 1, \pm 1, \pm 1, \pm 1)$, together with the eight points $(\pm 2, 0, 0, 0)$, $(0, \pm 2, 0, 0)$, $(0, 0, \pm 2, 0)$, and $(0, 0, 0, \pm 2)$.

The theorem that these are all the regular polytopes is significantly harder to prove than the result sketched above for three dimensions. The complete list was obtained by Schafli in the mid nineteenth century; the first proof that there are no others was given by Donald Coxeter in 1969.

We therefore know that the regular polytopes in dimensions three and higher fall into three families—the n -dimensional versions of the tetrahedron, cube, and octahedron—together with five “exceptional” examples—the dodecahedron, the icosahedron, and the three four-dimensional polytopes just described. This situation is typical of many classification theorems. The exceptional examples, often called “sporadic,” tend to have a very high degree of symmetry—it is almost as if we have no right to expect this degree of symmetry to be possible, but just occasionally by a happy chance it is. The families and sporadic examples that occur in different classification results are often closely related, and this can be a sign of deep connections between areas that do not at first appear to be connected at all.

Sometimes one does not try to classify all mathematical structures of a given kind, but instead identifies a certain class of “basic” structures out of which all the others can be built in a simple way. A good analogy for this is the set of primes, out of which all other integers can be built as products. Finite groups, for example, are all “products” of certain basic groups that are called *simple*. THE CLASSIFICATION OF FINITE SIMPLE GROUPS [V.8], one of the most famous theorems of twentieth-century mathematics, is discussed in part V.

For more on this style of classification theorem, see also LIE THEORY [III.50].

2.2 Equivalence, Nonequivalence, and Invariants

There are many situations in mathematics where two objects are, strictly speaking, different, but where we are not interested in the difference. In such situations we want to regard the objects as “essentially the same,” or “equivalent.” Equivalence of this kind is expressed formally by the notion of an EQUIVALENCE RELATION [I.2 §2.3].

For example, a topologist regards two shapes as essentially the same if one is a continuous deformation of

the other, as we saw in [I.3 §6.4]. As pointed out there, a sphere is the same as a cube in this sense, and one can also see that the surface of a doughnut, that is, a torus, is essentially the same as the surface of a teacup. (To turn the teacup into a doughnut, let the handle expand while the cup part is gradually swallowed up into it.) It is equally obvious, intuitively speaking, that a sphere is *not* essentially the same as a torus, but this is much harder to prove.

Why should nonequivalence be harder to prove than equivalence? The answer is that in order to show that two objects are equivalent, all one has to do is find a single transformation that demonstrates this equivalence. However, to show that two objects are not equivalent, one must somehow consider *all possible* transformations and show that not one of them works. How can one rule out the existence of some wildly complicated continuous deformation that is impossible to visualize but happens, remarkably, to turn a sphere into a torus?

Here is a sketch of a proof. The sphere and the torus are examples of *compact orientable surfaces*, which means, roughly speaking, two-dimensional shapes that occupy a finite portion of space and have no boundary. Given any such surface, one can find an equivalent surface that is built out of triangles and is topologically the same. Here is a famous theorem of EULER [VI.18].

Let P be a polyhedron that is topologically the same as a sphere, and suppose that it has V vertices, E edges, and F faces. Then $V - E + F = 2$.

For example, if P is an icosahedron, then it has twelve vertices, thirty edges, and twenty faces, and $12 - 30 + 20$ is indeed equal to 2.

For this theorem, it is not in fact important that the triangles are flat: we can draw them on the original sphere, except that now they are spherical triangles. It is just as easy to count vertices, edges, and faces when we do this, and the theorem is still valid. A network of triangles drawn on a sphere is called a *triangulation* of the sphere.

Euler's theorem tells us that $V - E + F = 2$ regardless of what triangulation of the sphere we take. Moreover, the formula is still valid if the surface we triangulate is not a sphere but another shape that is topologically equivalent to the sphere, since triangulations can be continuously deformed without V , E , or F changing.

More generally, one can triangulate *any* surface, and evaluate $V - E + F$. The result is called the *Euler number* of that surface. For this definition to make sense, we need the following fact, which is a generalization of Euler's theorem (and which is not much harder to prove than the original result).

- (i) *Although a surface can be triangulated in many ways, the quantity $V - E + F$ will be the same for all triangulations.*

If we continuously deform the surface and continuously deform one of its triangulations at the same time, we can deduce that the Euler number of the new surface is the same as that of the old one. In other words, fact (i) above has the following interesting consequence.

- (ii) *If two surfaces are continuous deformations of each other, then they have the same Euler number.*

This gives us a potential method for showing that surfaces are not equivalent: if they have different Euler numbers then we know from the above that they are not continuous deformations of each other. The Euler number of the torus turns out to be 0 (as one can show by calculating $V - E + F$ for any triangulation), and that completes the proof that the sphere and the torus are not equivalent.

The Euler number is an example of an *invariant*. This means a function ϕ , the domain of which is the set of all objects of the kind one is studying, with the property that if X and Y are equivalent objects, then $\phi(X) = \phi(Y)$. To show that X is not equivalent to Y , it is enough to find an invariant ϕ for which $\phi(X)$ and $\phi(Y)$ are different. Sometimes the values ϕ takes are numbers (as with the Euler number), but often they will be more complicated objects such as polynomials or groups.

It is perfectly possible for $\phi(X)$ to equal $\phi(Y)$ even when X and Y are not equivalent. An extreme example would be the invariant ϕ that simply took the value 0 for every object X . However, sometimes it is so hard to prove that objects are not equivalent that invariants can be considered useful and interesting even when they work only part of the time.

There are two main properties that one looks for in an invariant ϕ , and they tend to pull in opposite directions. One is that it should be as *fine* as possible: that is, as often as possible $\phi(X)$ and $\phi(Y)$ are different if X and Y are not equivalent. The other is that as often as possible one should actually be able to establish when $\phi(X)$ is different from $\phi(Y)$. There is not much use in having a fine invariant if it is impossible to calculate. (An extreme example would be the "trivial" invariant that simply mapped each X to its equivalence class. It is as fine as possible, but unless we have some independent means of specifying it, then it does not represent an advance on the original problem of showing that two objects are not equivalent.) The most powerful invari-

ants therefore tend to be ones that can be calculated, but not very easily.

In the case of compact orientable surfaces, we are lucky: not only is the Euler number an invariant that is easy to calculate, but it also classifies the compact orientable surfaces completely. To be precise, k is the Euler number of a compact orientable surface if and only if it is of the form $2 - 2g$ for some nonnegative integer g (so the possible Euler numbers are $2, 0, -2, -4, \dots$), and two compact orientable surfaces with the same Euler number are equivalent. Thus, if we regard equivalent surfaces as the same, then the number g gives us a complete specification of a surface. It is called the *genus* of the surface, and can be interpreted geometrically as the number of “holes” the surface has (so the genus of the sphere is 0 and that of the torus is 1).

For other examples of invariants, see ALGEBRAIC TOPOLOGY [IV.10] and KNOT POLYNOMIALS [III.46].

3 Generalizing

When an important mathematical definition is formulated, or theorem proved, that is rarely the end of the story. However clear a piece of mathematics may seem, it is nearly always possible to understand it better, and one of the most common ways of doing so is to present it as a special case of something more general. There are various different kinds of generalization, of which we discuss a few here.

3.1 Weakening Hypotheses and Strengthening Conclusions

The number 1729 is famous for being expressible as the sum of two cubes in two different ways: it is $1^3 + 12^3$ and also $9^3 + 10^3$. Let us now try to decide whether there is a number that can be written as the sum of four cubes in ten different ways.

At first this problem seems alarmingly difficult. It is clear that any such number, if it exists, must be very large and would be extremely tedious to find if we simply tested one number after another. So what can we do that is better than this?

The answer turns out to be that we should weaken our hypotheses. The problem we wish to solve is of the following general kind. We are given a sequence a_1, a_2, a_3, \dots of positive integers and we are told that it has a certain property. We must then prove that there is a positive integer that can be written as a sum of four terms of the sequence in ten different ways. This is perhaps an artificial way of thinking about the problem since the property we assume of the sequence is the

property of “being the sequence of cubes,” which is so specific that it is more natural to think of it as an *identification* of the sequence. However, this way of thinking encourages us to consider the possibility that the conclusion might be true for a much wider class of sequences. And indeed this turns out to be the case.

There are a thousand cubes less than or equal to 1 000 000 000. We shall now see that this property alone is sufficient to guarantee that there is a number that can be written as the sum of four cubes in ten different ways. That is, if a_1, a_2, a_3, \dots is *any* sequence of positive integers, and if none of the first thousand terms exceeds 1 000 000 000, then some number can be written as the sum of four terms of the sequence in ten different ways.

To prove this, all we have to do is notice that the number of different ways of choosing four distinct terms from the sequence $a_1, a_2, \dots, a_{1000}$ is $1000 \times 999 \times 998 \times 997 / 24$, which is greater than $40 \times 1\,000\,000\,000$. The sum of any four terms of the sequence cannot exceed $4 \times 1\,000\,000\,000$. It follows that the average number of ways of writing one of the first 4 000 000 000 numbers as the sum of four terms of the sequence is at least ten. But if the average number of representations is at least ten, then there must certainly be numbers that have at least this number of representations.

Why did it help to generalize the problem in this way? One might think that it would be harder to prove a result if one assumed less. However, that is often not true. The less you assume, the fewer options you have when trying to use your assumptions, and that can speed up the search for a proof. Had we not generalized the problem above, we would have had too many options. For instance, we might have found ourselves trying to solve very difficult Diophantine equations involving cubes rather than noticing the easy counting argument. In a way, it was only once we had weakened our hypotheses that we understood the true nature of the problem.

We could also think of the above generalization as a strengthening of the conclusion: the problem asks for a statement about cubes, and we prove not just that but much more besides. There is no clear distinction between weakening hypotheses and strengthening conclusions, since if we are asked to prove a statement of the form $P \Rightarrow Q$, we can always reformulate it as $\neg Q \Rightarrow \neg P$. Then, if we weaken P we are weakening the hypotheses of $P \Rightarrow Q$ but strengthening the conclusion of $\neg Q \Rightarrow \neg P$.

3.2 Proving a More Abstract Result

A famous result in modular arithmetic, known as *Fermat's little theorem* (see MODULAR ARITHMETIC [III.60]),

states that if p is a prime and a is not a multiple of p , then a^{p-1} leaves a remainder of 1 when you divide by p . That is, a^{p-1} is congruent to 1 mod p .

There are several proofs of this result, one of which is a good illustration of a certain kind of generalization. Here is the argument in outline. The first step is to show that the numbers $1, 2, \dots, p-1$ form a GROUP [I.3 §2.1] under multiplication mod p . (This means multiplication followed by taking the remainder on division by p . For example, if $p = 7$ then the “product” of 3 and 6 is 4, since 4 is the remainder when you divide 18 by 7.) The next step is to note that if $1 \leq a \leq p-1$ then the powers of $a \pmod{p}$ form a subgroup of this group. Moreover, the size of the subgroup is the smallest positive integer m such that a^m is congruent to 1 mod p . One then applies *Lagrange’s theorem*, which states that the size of a group is always divisible by the size of any of its subgroups. In this case, the size of the group is $p-1$, from which it follows that $p-1$ is divisible by m . But then, since $a^m = 1$, it follows that $a^{p-1} = 1$.

This argument shows that Fermat’s little theorem is, when viewed appropriately, just one special case of Lagrange’s theorem. (The word “just” is, however, a little misleading, because it is not wholly obvious that the integers mod p form a group in the way stated. This fact is proved using EUCLID’S ALGORITHM [III.22].)

Fermat could not have viewed his theorem in this way, since the concept of a group had not been invented when he proved it. Thus, the abstract concept of a group helps one to see Fermat’s little theorem in a completely new way: it can be viewed as a special case of a more general result, but a result that cannot even be stated until one has developed some new, abstract concepts.

This process of abstraction has many benefits. Most obviously, it provides us with a more general theorem, one that has many other interesting particular cases. Once we see this, then we can prove the general result once and for all rather than having to prove each case separately. A related benefit is that it enables us to see connections between results that may originally have seemed quite different. And finding surprising connections between different areas of mathematics almost always leads to significant advances in the subject.

3.3 Identifying Characteristic Properties

There is a marked contrast between the way one defines $\sqrt{2}$ and the way one defines $\sqrt{-1}$, or i as it is usually written. In the former case one begins, if one is being careful, by proving that there is exactly one positive real number that squares to 2. Then $\sqrt{2}$ is defined to be this number.

This style of definition is impossible for i since there is no real number that squares to -1 . So instead one asks the following question: if there were a number that squared to -1 , what could one say about it? Such a number would not be a real number, but that does not rule out the possibility of *extending* the real number system to a larger system that contains a square root of -1 .

At first it may seem as though we know precisely one thing about i : that $i^2 = -1$. But if we assume in addition that i obeys the normal rules of arithmetic, then we can do more interesting calculations, such as

$$(i+1)^2 = i^2 + 2i + 1 = -1 + 2i + 1 = 2i,$$

which implies that $(i+1)/\sqrt{2}$ is a square root of i .

From these two simple assumptions—that $i^2 = -1$ and that i obeys the usual rules of arithmetic—we can develop the entire theory of COMPLEX NUMBERS [I.3 §1.5] without ever having to worry about what i actually is. And in fact, once you stop to think about it, the existence of $\sqrt{2}$, though reassuring, is not in practice anything like as important as *its* defining properties, which are very similar to those of i : it squares to 2 and obeys the usual rules of arithmetic.

Many important mathematical generalizations work in a similar way. Another example is the definition of x^a when x and a are real numbers with x positive. It is difficult to make sense of this expression in a direct way unless a is a positive integer, and yet mathematicians are completely comfortable with it, whatever the value of a . How can this be? The answer is that what really matters about x^a is not its numerical value but its *characteristic properties* when one thinks of it as a function of a . The most important of these is the property that $x^{a+b} = x^a x^b$. Together with a couple of other simple properties, this completely determines the function x^a . More importantly, it is these characteristic properties that one uses when reasoning about x^a . This example is discussed in more detail in THE EXPONENTIAL AND LOGARITHMIC FUNCTIONS [III.25].

There is an interesting relationship between abstraction and classification. The word “abstract” is often used to refer to a part of mathematics where it is more common to use characteristic properties of an object than it is to argue directly from a definition of the object itself (though, as the example of $\sqrt{2}$ shows, this distinction can be somewhat hazy). The ultimate in abstraction is to explore the consequences of a system of axioms, such as those for a group or a vector space. However, sometimes, in order to reason about such algebraic structures, it is very helpful to classify them, and the result of classification is to make them more concrete again. For instance,

every finite-dimensional real vector space V is isomorphic to \mathbb{R}^n for some nonnegative integer n , and it is sometimes helpful to think of V as the concrete object \mathbb{R}^n , rather than as an algebraic structure that satisfies certain axioms. Thus, in a certain sense, classification is the opposite of abstraction.

3.4 Generalization after Reformulation

Dimension is a mathematical idea that is also a familiar part of everyday language: for example, we say that a photograph of a chair is a two-dimensional representation of a three-dimensional object, because the chair has height, breadth, and depth, but the image just has height and breadth. Roughly speaking, the dimension of a shape is the number of independent directions one can move about in while staying inside the shape, and this rough conception can be made mathematically precise (using the notion of a VECTOR SPACE [I.3 §2.3]).

If we are given any shape, then its dimension, as one would normally understand it, must be a nonnegative integer: it does not make much sense to say that one can move about in 1.4 independent directions, for example. And yet there is a rigorous mathematical theory of *fractional* dimension, in which for every nonnegative real number d you can find many shapes of dimension d .

How do mathematicians achieve the seemingly impossible? The answer is that they *reformulate* the concept of dimension and only then do they generalize it. What this means is that they give a new definition of dimension with the following two properties.

- (i) For all “simple” shapes the new definition agrees with the old one. For example, under the new definition a line will still be one dimensional, a square two dimensional, and a cube three dimensional.
- (ii) With the new definition it is no longer obvious that the dimension of every shape must be a positive integer.

There are several ways of doing this, but most of them focus on the differences between length, area, and volume. Notice that a line segment of length 2 can be expressed as a union of two nonoverlapping line segments of length 1, a square of side-length 2 can be expressed as a union of four nonoverlapping squares of side-length 1, and a cube of side-length 2 can be expressed as a union of eight nonoverlapping cubes of side-length 1. It is because of this that if you enlarge a d -dimensional shape by a factor r , then its d -dimensional “volume” is multiplied by r^d . Now suppose that you would like to exhibit a shape of dimension 1.4. One way

of doing it is to let $r = 2^{5/7}$, so that $r^{1.4} = 2$, and find a shape X such that if you expand X by a factor of r , then the expanded shape can be expressed as a union of two disjoint copies of X . Two copies of X ought to have twice the “volume” of X itself, so the dimension d of X ought to satisfy the equation $r^d = 2$. By our choice of r , this tells us that the dimension of X is 1.4. For more details, see DIMENSION [III.17].

Another concept that seems at first to make no sense is *noncommutative geometry*. The word “commutative” applies to BINARY OPERATIONS [I.2 §2.4] and therefore belongs to algebra rather than geometry, so what could “noncommutative geometry” possibly mean?

By now the answer should not be a surprise: one reformulates part of geometry in terms of a certain algebraic structure and then generalizes the algebra. The algebraic structure involves a commutative binary operation, so one can generalize the algebra by allowing the binary operation not to be commutative.

The part of geometry in question is the study of MANIFOLDS [I.3 §6.9]. Associated with a manifold X is the set $C(X)$ of all continuous complex-valued functions defined on X . Given two functions f, g in $C(X)$, and two complex numbers λ and μ , the linear combination $\lambda f + \mu g$ is another continuous complex-valued function, so it also belongs to $C(X)$. Therefore, $C(X)$ is a vector space. However, one can also *multiply* f and g to form the continuous function fg (defined by $(fg)(x) = f(x)g(x)$). This multiplication has various natural properties (for instance, $f(g+h) = fg + fh$ for all functions f, g , and h) that make $C(X)$ into an *algebra*, and even a C^* -ALGEBRA [IV.19 §3]. It turns out that a great deal of the geometry of a compact manifold X can be reformulated purely in terms of the corresponding C^* -algebra $C(X)$. The word “purely” here means that it is not necessary to refer to the manifold X in terms of which the algebra $C(X)$ was originally defined—all one uses is the fact that $C(X)$ is an algebra. This raises the possibility that there might be algebras that do *not* arise geometrically, but to which the reformulated geometrical concepts nevertheless apply.

An algebra has two binary operations: addition and multiplication. Addition is always assumed to be commutative, but multiplication is not: when multiplication is commutative as well, one says that the algebra is commutative. Since fg and gf are clearly the same function, the algebra $C(X)$ is a commutative C^* -algebra, so the algebras that arise geometrically are always commutative. However, many geometrical concepts, once they have been reformulated in algebraic terms, continue to make sense for noncommutative C^* -algebras, and that

is why the phrase “noncommutative” geometry is used. For more details, see OPERATOR ALGEBRAS [IV.19 §5].

This process of reformulating and then generalizing underlies many of the most important advances in mathematics. Let us briefly look at a third example. THE FUNDAMENTAL THEOREM OF ARITHMETIC [V.16] is, as its name suggests, one of the foundation stones of number theory: it states that every positive integer can be written in exactly one way as a product of prime numbers. However, number theorists like to look at enlarged number systems, and for most of these the obvious analogue of the fundamental theorem of arithmetic is no longer true. For example, in the RING [III.82 §1] of numbers of the form $a + b\sqrt{-5}$ (where a and b are required to be integers), the number 6 can be written either as 2×3 or as $(1 + \sqrt{-5}) \times (1 - \sqrt{-5})$. Since none of the numbers 2, 3, $1 + \sqrt{-5}$, or $1 - \sqrt{-5}$ can be decomposed further, the number 6 has two genuinely different prime factorizations in this ring.

There is, however, a natural way of generalizing the concept of “number” to include IDEAL NUMBERS [III.82 §2] that allow one to prove a version of the fundamental theorem of arithmetic in rings such as the one just defined. First, we must reformulate: we associate with each number y the set of all its multiples δy , where δ belongs to the ring. This set, which is denoted (y) , has the following closure property: if α and β belong to (y) and δ and ϵ are any two elements of the ring, then $\delta\alpha + \epsilon\beta$ belongs to (y) .

A subset of a ring with that closure property is called an *ideal*. If the ideal is of the form (y) for some number y , then it is called a *principal ideal*. However, there are ideals that are not principal, so we can think of the set of ideals as generalizing the set of elements of the original ring (once we have reformulated each element y as the principal ideal (y)). It turns out that there are natural notions of addition and multiplication that can be applied to ideals. Moreover, it makes sense to define an ideal I to be “prime” if the only way of writing I as a product JK is if one of J and K is a “unit.” In this enlarged set, unique factorization turns out to hold. These concepts give us a very useful way to measure “the extent to which unique factorization fails” in the original ring. For more details, see ALGEBRAIC NUMBERS [IV.3 §7].

3.5 Higher Dimensions and Several Variables

We have already seen that the study of polynomial equations becomes much more complicated when one looks not just at single equations in one variable, but at systems of equations in several variables. Similarly, we have

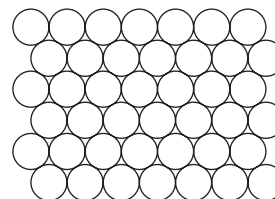


Figure 1 The densest possible packing of circles in the plane.

seen that PARTIAL DIFFERENTIAL EQUATIONS [I.3 §5.4], which can be thought of as differential equations involving several variables, are typically much more difficult to analyze than ordinary differential equations, that is, differential equations in just one variable. These are two notable examples of a process that has generated many of the most important problems and results in mathematics, particularly over the last century or so: the process of generalization from one variable to several variables.

Suppose one has an equation that involves three real variables, x , y , and z . It is often useful to think of the triple (x, y, z) as an object in its own right, rather than as a collection of three numbers. Furthermore, this object has a natural interpretation: it represents a point in three-dimensional space. This geometrical interpretation is important, and goes a long way toward explaining why extensions of definitions and theorems from one variable to several variables are so interesting. If we generalize a piece of algebra from one variable to several variables, we can also think of what we are doing as generalizing from a one-dimensional setting to a higher-dimensional setting. This idea leads to many links between algebra and geometry, allowing techniques from one area to be used to great effect in the other.

4 Discovering Patterns

Suppose that you wish to fill the plane as densely as possible with nonoverlapping circles of radius 1. How should you do it? This question is an example of a so-called *packing problem*. The answer is known, and it is what one might expect: you should arrange the circles so that their centers form a triangular lattice, as shown in figure 1. In three dimensions a similar result is true, but much harder to prove: until recently it was a famous open problem known as the Kepler conjecture. Several mathematicians wrongly claimed to have solved it, but in 1998 a long and complicated solution, obtained with the help of a computer, was announced by Thomas Hales,

and although his solution has proved very hard to check, the consensus is that it is probably correct.

Questions about packing of spheres can be asked in any number of dimensions, but they become harder and harder as the dimension increases. Indeed, it is likely that the best density for a ninety-seven-dimensional packing, say, will never be known. Experience with similar problems suggests that the best arrangement will almost certainly not have a simple structure such as one sees in two dimensions, so that the only method for finding it would be a “brute-force search” of some kind. However, to search for the best possible complicated structure is not feasible: even if one could somehow reduce the search to finitely many possibilities, there would be far more of them than one could feasibly check.

When a problem looks too difficult to solve, one should not give up completely. A much more productive reaction is to formulate related but more approachable questions. In this case, instead of trying to discover the very best packing, one can simply see how dense a packing one can find. Here is a sketch of an argument that gives a goodish packing in n dimensions, when n is large. One begins by taking a *maximal packing*: that is, one simply picks sphere after sphere until it is no longer possible to pick another one without it overlapping one of the spheres already chosen. This means that, for at least one of the spheres we have chosen, the distance from its center to x is less than 2—otherwise we could take a unit sphere about x and it would not overlap any of the other spheres. Therefore, if we take all the spheres in the collection and expand them by a factor of 2, then we cover all of \mathbb{R}^n . Since expanding an n -dimensional sphere by a factor of 2 increases its (n -dimensional) volume by a factor of 2^n , the proportion of \mathbb{R}^n covered by the unexpanded spheres must be at least 2^{-n} .

Notice that in the above argument we learned nothing at all about the nature of the arrangements of spheres with density 2^{-n} . All we did was take a maximal packing, and that can be done in a very haphazard way. This is in marked contrast with the approach that worked in two dimensions, where we defined a specific pattern of circles.

This contrast pervades all of mathematics. For some problems, the best approach is to build a highly structured pattern that does what you want, while for others—usually problems for which there is no hope of obtaining an exact answer—it is better to look for less specific arrangements. “Highly structured” in this context often means “possessing a high degree of symmetry.”

The triangular lattice is a rather simple pattern, but some highly structured patterns are much more complicated, and much more of a surprise when they are discovered. A notable example occurs in packing problems. By and large, the higher the dimension you are working in, the more difficult it is to find good patterns, but an exception to this general rule occurs at twenty-four dimensions. Here, there is a remarkable construction, known as the *Leech lattice*, which gives rise to a miraculously dense packing. Formally, a *lattice* in \mathbb{R}^n is a subset Λ with the following three properties.

- (i) If x and y belong to Λ , then so do $x + y$ and $x - y$.
- (ii) If x belongs to Λ , then x is *isolated*. That is, there is some $d > 0$ such that the distance between x and any other point of Λ is at least d .
- (iii) Λ is not contained in any $(n - 1)$ -dimensional subspace of \mathbb{R}^n .

A good example of a lattice is the set \mathbb{Z}^n of all points in \mathbb{R}^n with integer coordinates. If one is searching for a dense packing, then it is a good idea to look at lattices, since if you know that every nonzero point in a lattice has distance at least d from 0, then you know that *any* two points have distance at least d from each other. This is because the distance between x and y is the same as the distance between 0 and $y - x$, both of which lie in the lattice if x and y do. Thus, instead of having to look at the whole lattice, one can get away with looking at a small portion around 0.

In twenty-four dimensions it can be shown that there is a lattice Λ with the following additional properties, and that it is unique, in the sense that any other lattice with those properties is just a rotation of the first one.

- (iv) There is a 24×24 matrix M with DETERMINANT [III.15] equal to 1 such that Λ consists of all integer combinations of the columns of M .
- (v) If v is a point in Λ , then the square of the distance from 0 to v is an even integer.
- (vi) The nearest nonzero vector to 0 is at distance 2. Thus, the balls of radius 1 about the points in Λ form a packing of \mathbb{R}^{24} .

The nearest nonzero vector is far from unique: in fact there are 196 560 of them, which is a remarkably large number considering that these points must all be at distance at least 2 from each other.

The Leech lattice also has an extraordinary degree of symmetry. To be precise, it has 8 315 553 613 086 720 000 rotational symmetries. (This number equals $2^{22} \cdot 3^9 \cdot 5^4 \cdot 7^2 \cdot 11 \cdot 13 \cdot 23$.) If you take the QUOTIENT [I.3 §3.3]

of its symmetry group by the subgroup consisting of the identity and minus the identity, then you obtain the *Conway group* Co_1 , which is one of the famous sporadic SIMPLE GROUPS [V.8]. The existence of so many symmetries makes it easier still to determine the smallest distance from 0 of any nonzero point of the lattice, since once you have checked one distance you have automatically checked lots of others (just as, in the triangular lattice, the six-fold rotational symmetry tells us that the distances from 0 to its six neighbors are all the same).

These facts about the Leech lattice illustrate a general principle of mathematical research: often, if a mathematical construction has one remarkable property, it will have others as well. In particular, a high degree of symmetry will often be related to other interesting features. So, although it is a surprise that the Leech lattice exists at all, it is not as surprising when one then discovers that it gives an extremely dense packing of \mathbb{R}^{24} . In fact, it was shown in 2004 by Henry Cohn and Abhinav Kumar that it gives the densest possible packing of spheres in twenty-four-dimensional space, at least among all packings derived from lattices. It is probably the densest packing of any kind, but this has not yet been proved.

5 Explaining Apparent Coincidences

The largest of all the sporadic finite simple groups is called the *Monster group*. Its name is partly explained by the fact that it has $2^{46} \cdot 3^{20} \cdot 5^9 \cdot 7^6 \cdot 11^2 \cdot 13^3 \cdot 17 \cdot 19 \cdot 23 \cdot 29 \cdot 31 \cdot 41 \cdot 47 \cdot 59 \cdot 71$ elements. How can one hope to understand a group of this size?

One of the best ways is to show that it is a group of symmetries of some other mathematical object (see the article on REPRESENTATION THEORY [IV.12] for much more on this theme), and the smaller that object is, the better. We have just seen that another large sporadic group, the Conway group Co_1 , is closely related to the symmetry group of the Leech lattice. Might there be a lattice that played a similar role for the Monster group?

It is not hard to show that there will be at least *some* lattice that works, but more challenging is to find one of small dimension. It has been shown that the smallest possible dimension that can be used is 196 883.

Now let us turn to a different branch of mathematics. If you look at the article about ALGEBRAIC NUMBERS [IV.3 §8] you will see a definition of a function $j(z)$, called the *elliptic modular function*, of central importance in algebraic number theory. It is given as the sum

of a series that starts

$$j(z) = e^{-2\pi iz} + 744 + 196\,884e^{2\pi iz} \\ + 21\,493\,760e^{4\pi iz} + 864\,299\,970e^{6\pi iz} + \dots$$

Rather intriguingly, the coefficient of $e^{2\pi iz}$ in this series is 196 884, one more than the smallest possible dimension of a lattice that has the Monster group as its group of symmetries.

It is not obvious how seriously we should take this observation, and when it was first made by John McKay opinions differed about it. Some believed that it was probably just a coincidence, since the two areas seemed to be so different and unconnected. Others took the attitude that the function $j(z)$ and the Monster group are so important in their respective areas, and the number 196 883 so large, that the surprising numerical fact was probably pointing to a deep connection that had not yet been uncovered.

It turned out that the second view was correct. After studying the coefficients in the series for $j(z)$, McKay and John Thompson were led to a conjecture that related them all (and not just 196 884) to the Monster group. This conjecture was extended by John Conway and Simon Norton, who formulated the “Monstrous moonshine” conjecture, which was eventually proved by Richard Borcherds in 1992. (The word “moonshine” reflects the initial disbelief that there would be a serious relationship between the Monster group and the j -function.)

In order to prove the conjecture, Borcherds introduced a new algebraic structure, which he called a VERTEX ALGEBRA [IV.13]. And to analyze vertex algebras, he used results from STRING THEORY [IV.13 §2]. In other words, he explained the connection between two very different-looking areas of pure mathematics with the help of concepts from theoretical physics.

This example demonstrates in an extreme way another general principle of mathematical research: if you can obtain the same series of numbers (or the same structure of a more general kind) from two different mathematical sources, then those sources are probably not as different as they seem. Moreover, if you can find one deep connection, you will probably be led to others. There are many other examples where two completely different calculations give the same answer, and many of them remain unexplained. This phenomenon results in some of the most difficult and fascinating unsolved problems in mathematics. (See the introduction to MIRROR SYMMETRY [IV.14] for another example.)

Interestingly, the j -function leads to a second famous mathematical “coincidence.” There may not seem to be

anything special about the number $e^{\pi\sqrt{163}}$, but here is the beginning of its decimal expansion:

$$e^{\pi\sqrt{163}} = 262\,537\,412\,640\,768\,743.99999999999925\dots,$$

which is astonishingly close to an integer. Again it is initially tempting to dismiss this as a coincidence, but one should think twice before yielding to the temptation. After all, there are not all that many numbers that can be defined as simply as $e^{\pi\sqrt{163}}$, and each one has a probability of less than one in a million million of being as close to an integer as $e^{\pi\sqrt{163}}$ is. In fact, it is not a coincidence at all: for an explanation see ALGEBRAIC NUMBERS [IV.3 §8].

6 Counting and Measuring

How many rotational symmetries are there of a regular icosahedron? Here is one way to work it out. Choose a vertex v of the icosahedron and let v' be one of its neighbors. An icosahedron has twelve vertices, so there are twelve places where v could end up after the rotation. Once we know where v goes, there are five possibilities for v' (since each vertex has five neighbors and v' must still be a neighbor of v after the rotation). Once we have determined where v and v' go, there is no further choice we can make, so the number of rotational symmetries is $5 \times 12 = 60$.

This is a simple example of a *counting argument*, that is, an answer to a question that begins “How many.” However, the word “argument” is at least as important as the word “counting,” since we do not put all the symmetries in a row and say “one, two, three, . . . , sixty,” as we might if we were counting in real life. What we do instead is come up with a reason for the number of rotational symmetries being 5×12 . At the end of the process, we understand more about those symmetries than merely how many there are. Indeed, it is possible to go further and show that the group of rotations of the icosahedron is A_5 , the ALTERNATING GROUP [III.70] on five elements.

6.1 Exact Counting

Here is a more sophisticated counting problem. A *one-dimensional random walk* of n steps is a sequence of integers $a_0, a_1, a_2, \dots, a_n$, such that for each i the difference $a_i - a_{i-1}$ is either 1 or -1 . For example, $0, 1, 2, 1, 2, 1, 0, -1$ is a seven-step random walk. The number of n -step random walks that start at 0 is clearly 2^n , since there are two choices for each step (either you add 1 or you subtract 1).

Now let us try a slightly harder problem. How many walks of length $2n$ are there that start *and end* at 0? (We look at walks of length $2n$ since a walk that starts and ends in the same place must have an even number of steps.)

In order to think about this problem, it helps to use the letters R and L (for “right” and “left”) to denote adding 1 and subtracting 1, respectively. This gives us an alternative notation for random walks that start at 0: for example, the walk $0, 1, 2, 1, 2, 1, 0, -1$ would be rewritten as RRLRLLL. Now a walk will end at 0 if and only if the number of Rs is equal to the number of Ls. Moreover, if we are told the set of steps where an R occurs, then we know the entire walk. So what we are counting is the number of ways of choosing n of the $2n$ steps as the steps where an R will occur. And this is well-known to be $(2n)!/(n!)^2$.

Now let us look at a related quantity that is considerably less easy to determine: the number $W(n)$ of walks of length $2n$ that start and end at 0 and are never negative. Here, in the notation introduced for the previous problem, is a list of all such walks of length 6: RRRLLL, RRLRLL, RRLRL, RLRRLL, and RLRLRL.

Now three of these five walks do not just start and end at 0 but visit it in the middle: RRLRLL visits it after four steps, RLRRLL after two, and RLRLRL after two and four. Suppose we have a walk of length $2n$ that is never negative and visits 0 for the first time after $2k$ steps. Then the remainder of the walk is a walk of length $2(n-k)$ that starts and ends at 0 and is never negative. There are $W(n-k)$ of these. As for the first $2k$ steps of such a walk, they must begin with R and end with L, and in between must never visit 0. This means that between the initial R and the final L they give a walk of length $2(k-1)$ that starts and ends at 1 and is never less than 1. The number of such walks is clearly the same as $W(k-1)$. Therefore, since the first visit to 0 must take place after $2k$ steps for some k between 1 and n , W satisfies the following slightly complicated recurrence relation:

$$W(n) = W(0)W(n-1) + \dots + W(n-1)W(0).$$

Here, $W(0)$ is taken to be equal to 1.

This allows us to calculate the first few values of W . We have $W(1) = W(0)W(0) = 1$, which is easier to see directly: the only possibility is RL. Then $W(2) = W(1)W(0) + W(0)W(1) = 2$, and $W(3)$, which counts the number of such walks of length 6, equals $W(0)W(2) + W(1)W(1) + W(2)W(0) = 5$, confirming our earlier calculation.

Of course, it would not be a good idea to use the recurrence relation directly if one wished to work out $W(n)$

for large values of n such as 10^{10} . However, the recurrence is of a sufficiently nice form that it is amenable to treatment by GENERATING FUNCTIONS [IV.22 §2.4], as is explained in ENUMERATIVE AND ALGEBRAIC COMBINATORICS [IV.22 §3]. (To see the connection with that discussion, replace the letters R and L by the square brackets [and], respectively. A legal bracketing then corresponds to a walk that is never negative.)

The argument above gives an efficient way of calculating $W(n)$ exactly. There are many other exact counting arguments in mathematics. Here is a small further sample of quantities that mathematicians know how to count exactly without resorting to “brute force.” (See the introduction to [IV.22] for a discussion of when one regards a counting problem as solved.)

(i) The number $r(n)$ of regions that a plane is cut into by n lines if no two of the lines are parallel and no three concurrent. The first four values of $r(n)$ are 2, 4, 7, and 11. It is not hard to prove that $r(n) = r(n-1) + n$, which leads to the formula $r(n) = \frac{1}{2}n(n+3)$. This statement, and its proof, can be generalized to higher dimensions.

(ii) The number $s(n)$ of ways of writing n as a sum of four squares. Here we allow zero and negative numbers and we count different orderings as different (so, for example, $1^2 + 3^2 + 4^2 + 2^2$, $3^2 + 4^2 + 1^2 + 2^2$, $1^2 + (-3)^2 + 4^2 + 2^2$, and $0^2 + 1^2 + 2^2 + 5^2$ are considered to be four different ways of writing 30 as a sum of four squares). It can be shown that $s(n)$ is equal to 8 times the sum of all the divisors of n that are not multiples of 4. For example, the divisors of 12 are 1, 2, 3, 4, 6, and 12, of which 1, 2, 3, and 6 are not multiples of 4. Therefore $s(12) = 8(1 + 2 + 3 + 6) = 96$. The different ways are $1^2 + 1^2 + 1^2 + 3^2$, $0^2 + 4^2 + 4^2 + 4^2$, and the other expressions that can be obtained from these ones by reordering and replacing positive integers by negative ones.

(iii) The number of lines in space that meet a given four lines L_1, L_2, L_3 , and L_4 when those four are in “general position.” (This means that they do not have special properties such as two of them being parallel or intersecting each other.) It turns out that for any *three* such lines, there is a subset of \mathbb{R}^3 known as a *quadric surface* that contains them, and that this quadric surface is unique. Let us take the surface for L_1, L_2 , and L_3 and call it S .

The surface S has some interesting properties that allow us to solve the problem. The main one is that one can find a continuous family of lines (that is, a collection of lines $L(t)$, one for each real number t , that varies continuously with t) that, between them, make up the surface S and include each of the lines L_1, L_2 , and L_3 .

But there is also *another* such continuous family of lines $M(s)$, each of which meets every line $L(t)$ in exactly one point. In particular, every line $M(s)$ meets all of L_1, L_2 , and L_3 , and in fact every line that meets all of L_1, L_2 , and L_3 must be one of the lines $M(s)$.

It can be shown that L_4 intersects the surface S in exactly two points, P and Q. Now P lies in some line $M(s)$ from the second family, and Q lies in some other line $M(s')$ (which must be different, or else L_4 would equal $M(s)$ and intersect L_1, L_2 , and L_3 , contradicting the fact that the lines L_i are in general position). Therefore, the two lines $M(s)$ and $M(s')$ intersect all four of the lines L_i . But every line that meets all the L_i has to be one of the lines $M(s)$ and has to go through either P or Q (since the lines $M(s)$ lie in S and L_4 meets S at only those two points). Therefore, the answer is 2.

This question can be generalized very considerably, and answered by means of a technique known as *Schubert calculus*.

(iv) The number $p(n)$ of ways of writing a positive integer n as a sum of smaller positive integers. When $n = 6$ this number is 11, since $6 = 1 + 1 + 1 + 1 + 1 + 1 = 2 + 1 + 1 + 1 + 1 = 2 + 2 + 1 + 1 = 2 + 2 + 2 = 3 + 1 + 1 + 1 = 3 + 2 + 1 = 3 + 3 = 4 + 1 + 1 = 4 + 2 = 5 + 1 = 6$. The function $p(n)$ is called the *partition function*. A remarkable formula, due to HARDY [VI.72] and RAMANUJAN [VI.81], gives an approximation $\alpha(n)$ to $p(n)$ that is so accurate that $p(n)$ is always the nearest integer to $\alpha(n)$.

6.2 Estimates

Once we have seen example (ii) above, it is natural to ask whether it can be generalized. Is there a formula for the number $t(n)$ of ways of writing n as a sum of ten sixth powers, for example? It is generally believed that the answer to this question is no, and certainly no such formula has been discovered. However, as with packing problems, even if an exact answer does not seem to be forthcoming, it is still very interesting to obtain *estimates*. In this case, one can try to define an easily calculated function f such that $f(n)$ is always *approximately* equal to $t(n)$. If even that is too hard, one can try to find *two* easily calculated functions L and U such that $L(n) \leq t(n) \leq U(n)$ for every n . If we succeed, then we call L a *lower bound* for t and U an *upper bound*. Here are a few examples of quantities that nobody knows how to count exactly, but for which there are interesting approximations, or at least interesting upper and lower bounds.

(i) Probably the most famous approximate counting problem in all of mathematics is to estimate $\pi(n)$, the number of prime numbers less than or equal to n . For small values of n , we can of course compute $\pi(n)$ exactly: for example, $\pi(20) = 8$ since the primes less than or equal to 20 are 2, 3, 5, 7, 11, 13, 17, and 19. However, there does not seem to be a useful formula for $\pi(n)$, and although it is easy to think of a brute-force algorithm for computing $\pi(n)$ —look at every number up to n , test whether it is prime, and keep count as you go along—such a procedure takes a prohibitively long time if n is at all large. Furthermore, it does not give us much insight into the nature of the function $\pi(n)$.

If, however, we modify the question slightly, and ask *roughly* how many primes there are up to n , then we find ourselves in the area known as ANALYTIC NUMBER THEORY [IV.4], a branch of mathematics with many fascinating results. In particular, the famous PRIME NUMBER THEOREM [V.33], proved by HADAMARD [VI.64] and DE LA VALLÉE POUSSIN [VI.66] at the end of the nineteenth century, states that $\pi(n)$ is approximately equal to $n/\log n$, in the sense that the ratio of $\pi(n)$ to $n/\log n$ converges to 1 as n tends to infinity.

This statement can be refined. It is believed that the “density” of primes close to n is about $1/\log n$, in the sense that a randomly chosen integer close to n has a probability of about $1/\log n$ of being prime. This would suggest that $\pi(n)$ should be about $\int_0^n dt/\log t$, a function of n that is known as the *logarithmic integral* of n , or $\text{li}(n)$.

How accurate is this estimate? Nobody knows, but THE RIEMANN HYPOTHESIS [V.33], perhaps the most famous unsolved problem in mathematics, is equivalent to the statement that $\pi(n)$ and $\text{li}(n)$ differ by at most $c\sqrt{n}\log n$ for some constant c . Since $\sqrt{n}\log n$ is much smaller than $\pi(n)$, this would tell us that $\text{li}(n)$ was an extremely good approximation to $\pi(n)$.

(ii) A *self-avoiding walk* of length n in the plane is a sequence of points $(a_0, b_0), (a_1, b_1), (a_2, b_2), \dots, (a_n, b_n)$ with the following properties.

- The numbers a_i and b_i are all integers.
- For each i , one obtains (a_i, b_i) from (a_{i-1}, b_{i-1}) by taking a horizontal or vertical step of length 1. That is, either $a_i = a_{i-1}$ and $b_i = b_{i-1} \pm 1$ or $a_i = a_{i-1} \pm 1$ and $b_i = b_{i-1}$.
- No two of the points (a_i, b_i) are equal.

The first two conditions tell us that the sequence forms a two-dimensional walk of length n , and the third says that this walk never visits any point more than once—hence the term “self-avoiding.”

Let $S(n)$ be the number of self-avoiding walks of length n that start at $(0, 0)$. There is no known formula for $S(n)$, and it is very unlikely that such a formula exists. However, quite a lot is known about the way the function $S(n)$ grows as n grows. For instance, it is fairly easy to prove that $S(n)^{1/n}$ converges to a limit c . The value of c is not known, but it has been shown (with the help of a computer) to lie between 2.62 and 2.68.

(iii) Let $C(t)$ be the number of points in the plane with integer coordinates contained in a circle of radius t about the origin. That is, $C(t)$ is the number of pairs (a, b) of integers such that $a^2 + b^2 \leq t^2$. A circle of radius t has area πt^2 , and the plane can be tiled by unit squares, each of which has a point with integer coordinates at its center. Therefore, when t is large it is fairly clear (and not hard to prove) that $C(t)$ is approximately πt^2 . However, it is much less clear how good this approximation is.

To make this question more precise, let us set $\epsilon(t)$ to equal $|C(t) - \pi t^2|$. That is, $\epsilon(t)$ is the *error* in πt^2 as an estimate for $C(t)$. It was shown in 1915, by Hardy and Landau, that $\epsilon(t)$ must be at least $c\sqrt{t}$ for some constant $c > 0$, and this estimate, or something very similar, probably gives the right order of magnitude for $\epsilon(t)$. However, the best upper bound, proved by Huxley in 1990 (the latest in a long line of successive improvements), is that $\epsilon(t)$ is at most $At^{46/73}$ for some constant A .

6.3 Averages

So far, our discussion of estimates and approximations has been confined to problems where the aim is to count mathematical objects of a given kind. However, that is by no means the only context in which estimates can be interesting. Given a set of objects, one may wish to know, besides its size, roughly what a typical one of those objects looks like. Many questions of this kind take the form of asking what the average value is of some numerical parameter that is associated with each object. Here are two examples.

(i) What is the average distance between the starting point and the endpoint of a self-avoiding walk of length n ? In this instance, the objects are self-avoiding walks of length n that start at $(0, 0)$, and the numerical parameter is the end-to-end distance.

Surprisingly, this is a notoriously difficult problem, and almost nothing is known. It is obvious that n is an upper bound for $S(n)$, but one would expect a typical self-avoiding walk to take many twists and turns and end up traveling much less far than n away from its starting

point. However, there is no known upper bound for $S(n)$ that is substantially better than n .

In the other direction, one would expect the end-to-end distance of a typical self-avoiding walk to be greater than that of an ordinary walk, to give it room to avoid itself. This would suggest that $S(n)$ is significantly greater than \sqrt{n} , but it has not even been proved that it is greater.

This is not the whole story, however, and the problem will be discussed further in section 8.

(ii) Let n be a large randomly chosen positive integer and let $\omega(n)$ be the number of distinct prime factors of n . On average, how large will $\omega(n)$ be? As it stands, this question does not quite make sense because there are infinitely many positive integers, so one cannot choose one randomly. However, one can make the question precise by specifying a large integer m and choosing a random integer n between m and $2m$. It then turns out that the average size of $\omega(n)$ is around $\log \log n$.

In fact, much more is known than this. If all you know about a RANDOM VARIABLE [III.73 §4] is its average, then a great deal of its behavior is not determined, so for many problems calculating averages is just the beginning of the story. In this case, Hardy and Ramanujan gave an estimate for the STANDARD DEVIATION [III.73 §4] of $\omega(n)$, showing that it is about $\sqrt{\log \log n}$. Then Erdős and Kac went even further and gave a precise estimate for the probability that $\omega(n)$ differs from $\log \log n$ by more than $c\sqrt{\log \log n}$, proving the surprising fact that the distribution of ω is approximately GAUSSIAN [III.73 §5].

To put these results in perspective, let us think about the range of possible values of $\omega(n)$. At one extreme, n might be a prime itself, in which case it obviously has just one prime factor. At the other extreme, we can write the primes in ascending order as p_1, p_2, p_3, \dots and take numbers of the form $n = p_1 p_2 \cdots p_k$. With the help of the prime number theorem, one can show that the order of magnitude of k is $\log n / \log \log n$, which is much bigger than $\log \log n$. However, the results above tell us that such numbers are exceptional: a typical number has a few distinct prime factors, but nothing like as many as $\log m / \log \log m$.

6.4 Extremal Problems

There are many problems in mathematics where one wishes to maximize or minimize some quantity in the presence of various constraints. These are called *extremal problems*. As with counting questions, there are some extremal problems for which one can realistically

hope to work out the answer exactly, and many more for which, even though an exact answer is out of the question, one can still aim to find interesting estimates. Here are some examples of both kinds.

(i) Let n be a positive integer and let X be a set with n elements. How many subsets of X can be chosen if none of these subsets is contained in any other?

A simple observation one can make is that if two different sets have the same size, then neither is contained in the other. Therefore, one way of satisfying the constraints of the problem is to choose all the sets of some particular size k . Now the number of subsets of X of size k is $n!/k!(n-k)!$, which is usually written $\binom{n}{k}$ (or nC_k), and the value of k for which $\binom{n}{k}$ is largest is easily shown to be $n/2$ if n is even and $(n+1)/2$ if n is odd. For simplicity let us concentrate on the case when n is even. What we have just proved is that it is possible to pick $\binom{n}{n/2}$ subsets of an n -element set in such a way that none of them contains any other. That is, $\binom{n}{n/2}$ is a lower bound for the problem. A result known as *Sperner's theorem* states that it is an upper bound as well. That is, if you choose *more* than $\binom{n}{n/2}$ subsets of X , then, however you do it, one of these subsets will be contained in another. Therefore, the question is answered exactly, and the answer is $\binom{n}{n/2}$. (When n is odd, then the answer is $\binom{n}{(n+1)/2}$, as one might now expect.)

(ii) Suppose that the two ends of a heavy chain are attached to two hooks on the ceiling and that the chain is not supported anywhere else. What shape will the hanging chain take?

At first, this question does not look like a maximization or minimization problem, but it can be quickly turned into one. That is because a general principle from physics tells us that the chain will settle in the shape that minimizes its potential energy. We therefore find ourselves asking a new question: let A and B be two points at distance d apart, and let \mathcal{C} be the set of all curves of length l that have A and B as their two endpoints. Which curve $C \in \mathcal{C}$ has the smallest potential energy? Here one takes the mass of any portion of the curve to be proportional to its length. The potential energy of the curve is equal to mgh , where m is the mass of the curve, g is the gravitational constant, and h is the height of the center of gravity of the curve. Since m and g do not change, another formulation of the question is: which curve $C \in \mathcal{C}$ has the smallest average height?

This problem can be solved by means of a technique known as *the calculus of variations*. Very roughly, the idea is this. We have a set, \mathcal{C} , and a function h defined on \mathcal{C} that takes each curve $C \in \mathcal{C}$ to its average height. We

are trying to minimize h , and a natural way to approach that task is to define some sort of derivative and look for a curve C at which this derivative is 0. Notice that the word “derivative” here does *not* refer to the rate of change of height as you move along the curve. Rather, it means the (linear) way that the average height of the entire curve changes in response to small perturbations of the curve. Using this kind of derivative to find a minimum is more complicated than looking for the stationary points of a function defined on \mathbb{R} , since \mathcal{C} is an infinite-dimensional set and is therefore much more complicated than \mathbb{R} . However, the approach can be made to work, and the curve that minimizes the average height is known. (It is called a *catenary*, after the Latin word for chain.) Thus, this is another minimization problem that has been answered exactly.

For a typical problem in the calculus of variations, one is trying to find a curve, or surface, or more general kind of function, for which a certain quantity is minimized or maximized. If a minimum or maximum exists (which is by no means automatic when one is working with an infinite-dimensional set, so this can be an interesting and important question), the object that achieves it satisfies a system of PARTIAL DIFFERENTIAL EQUATIONS [I.3 §5.4] known as the *Euler-Lagrange equations*. For more about this style of minimization or maximization, see VARIATIONAL METHODS [III.94] (and also OPTIMIZATION AND LAGRANGE MULTIPLIERS [III.66]).

(iii) How many numbers can you choose between 1 and n if no three of them are allowed to lie in an arithmetic progression? If $n = 9$ then the answer is 5. To see this, note first that no three of the five numbers 1, 2, 4, 8, 9 lie in an arithmetic progression. Now let us see if we can find six numbers that work.

If we make one of our numbers 5, then we must leave out either 4 or 6, or else we would have the progression 4, 5, 6. Similarly, we must leave out one of 3 and 7, one of 2 and 8, and one of 1 and 9. But then we have left out four numbers. It follows that we cannot choose 5 as one of the numbers.

We must leave out one of 1, 2, and 3, and one of 7, 8, and 9, so if we leave out 5 then we must include 4 and 6. But then we cannot include 2 or 8. But we must also leave out at least one of 1, 4, and 7, so we are forced to leave out at least four numbers.

An ugly case-by-case argument of this kind is feasible when $n = 9$, but as soon as n is at all large there are far too many cases for it to be possible to consider them all. For this problem, there does not seem to be a tidy answer that tells us exactly which is the largest set of integers between 1 and n that contains no arithmetic

progression of length 3. So instead one looks for upper and lower bounds on its size. To prove a lower bound, one must find a good way of constructing a large set that does not contain any arithmetic progressions, and to prove an upper bound one must show that *any* set of a certain size must necessarily contain an arithmetic progression. The best bounds to date are very far apart. In 1947, Behrend found a set of size $n/e^{c\sqrt{\log n}}$ that contains no arithmetic progression, and in 1999 Jean Bourgain proved that every set of size $Cn\sqrt{\log n/\log \log n}$ contains an arithmetic progression. (If it is not obvious to you that these numbers are far apart, then consider what happens when $n = 10^{100}$, say. Then $e^{c\sqrt{\log n}}$ is about 4 000 000, while $\sqrt{\log n/\log \log n}$ is about 6.5.)

(iv) Theoretical computer science provides many minimization problems: if one is programming a computer to perform a certain task, then one wants it to do so in as short a time as possible. Here is an elementary-sounding example: how many steps are needed to multiply two n -digit numbers together?

Even if one is not too precise about what is meant by a “step,” one can see that the traditional method, long multiplication, takes at least n^2 steps since, during the course of the calculation, each digit of the first number is multiplied by each digit of the second. One might imagine that this was necessary, but in fact there are clever ways of transforming the problem and dramatically reducing the time that a computer needs to perform a multiplication of this kind. The fastest known method uses THE FAST FOURIER TRANSFORM [III.26] to reduce the number of steps from n^2 to $Cn \log n \log \log n$. Since the logarithm of a number is much smaller than the number itself, one thinks of $Cn \log n \log \log n$ as being only just worse than a bound of the form Cn . Bounds of this form are called *linear*, and for a problem like this are clearly the best one can hope for, since it takes $2n$ steps even to read the digits of the two numbers.

Another question that is similar in spirit is whether there are fast algorithms for matrix multiplication. To multiply two $n \times n$ matrices using the obvious method one needs to do n^3 individual multiplications of the numbers in the matrices, but once again there are less obvious methods that do better. The main breakthrough on this problem was due to Strassen, who had the idea of splitting each matrix into four $n/2 \times n/2$ matrices and multiplying those together. At first it seems as though one has to calculate the products of eight pairs of $n/2 \times n/2$ matrices, but these products are related, and Strassen came up with *seven* such calculations from which the eight products could quickly be derived. One can then apply *recursion*: that is, use the same idea to

speed up the calculation of the seven $n/2 \times n/2$ matrix products, and so on.

Strassen's algorithm reduces the number of numerical multiplications from about n^3 to about $n^{\log_2 7}$. Since $\log_2 7$ is less than 2.81, this is a significant improvement, but only when n is large. His basic divide-and-conquer strategy has been developed further, and the current record is better than $n^{2.4}$. In the other direction, the situation is less satisfactory: nobody has found a proof that one needs to use significantly more than n^2 multiplications.

For more problems of a similar kind, see COMPUTATIONAL COMPLEXITY [IV.21] and THE MATHEMATICS OF ALGORITHM DESIGN [VII.5].

(v) Some minimization and maximization problems are of a more subtle kind. For example, suppose that one is trying to understand the nature of the differences between successive primes. The smallest such difference is 1 (the difference between 2 and 3), and it is not hard to prove that there is no largest difference (given any integer n greater than 1, none of the numbers between $n!+2$ and $n!+n$ is a prime). Therefore, there do not seem to be interesting maximization or minimization problems concerning these differences.

However, one can in fact formulate some fascinating problems if one first *normalizes* in an appropriate way. As was mentioned earlier in this section, the prime number theorem states that the density of primes near n is about $1/\log n$, so an average gap between two primes near n will be about $\log n$. If p and q are successive primes, we can therefore define a "normalized gap" to be $(q-p)/\log p$. The average value of this normalized gap will be 1, but is it sometimes much smaller and sometimes much bigger?

It was shown by Westzynthius in 1931 that even normalized gaps can be arbitrarily large, and it was widely believed that they could also be arbitrarily close to zero. (The famous twin-prime conjecture—that there are infinitely many primes p for which $p+2$ is also a prime—implies this immediately.) However, it took until 2005 for this to be proved, by Goldston, Pintz, and Yıldırım. (See ANALYTIC NUMBER THEORY [IV.4 §§6–8] for a discussion of this problem.)

7 Determining Whether Different Mathematical Properties Are Compatible

In order to understand a mathematical concept, such as that of a group or a manifold, there are various stages one typically goes through. Obviously it is a good

idea to begin by becoming familiar with a few representative examples of the structure, and also with techniques for building new examples out of old ones. It is also extremely important to understand the homomorphisms, or "structure-preserving functions," from one example of the structure to another, as was discussed in SOME FUNDAMENTAL MATHEMATICAL DEFINITIONS [I.3 §§4.1, 4.2].

Once one knows these basics, what is there left to understand? Well, for a general theory to be useful, it should tell us something about specific examples. For instance, as we saw in section 3.2, Lagrange's theorem can be used to prove Fermat's little theorem. Lagrange's theorem is a general fact about groups: that if G is a group of size n , then the size of any subgroup of G must be a factor of n . To obtain Fermat's little theorem, one applies Lagrange's theorem to the particular case when G is the multiplicative group of integers mod p . The conclusion one obtains—that a^p is always congruent to a —is far from obvious.

However, what if we want to know something about a group G that might not be true for all groups? That is, suppose that we wish to determine whether G has some property P that some groups have and others do not. Since we cannot prove that the property P follows from the group axioms, it might seem that we are forced to abandon the general theory of groups and look at the specific group G . However, in many situations there is an intermediate possibility: to identify some fairly general property Q that the group G has, and show that Q implies the more particular property P that interests us.

Here is an illustration of this sort of technique in a different context. Suppose we wish to determine whether the polynomial $p(x) = x^4 - 2x^3 - x^2 - 2x + 1$ has a real root. One method would be to study this particular polynomial and try to find a root. After quite a lot of effort we might discover that $p(x)$ can be factorized as $(x^2 + x + 1)(x^2 - 3x + 1)$. The first factor is always positive, but if we apply the quadratic formula to the second, we find that $p(x) = 0$ when $x = (3 \pm \sqrt{5})/2$. An alternative method, which uses a bit of general theory, is to notice that $p(1)$ is negative (in fact, it equals -3) and that $p(x)$ is large when x is large (because then the x^4 term is far bigger than anything else), and then to use the *intermediate value theorem*, the result that any continuous function that is negative somewhere and positive somewhere else must be zero somewhere in between.

Notice that, with the second approach, there was still some computation to do—finding a value of x for which $p(x)$ is negative—but that it was much easier than the computation in the first approach—finding a value of

x for which $p(x)$ is zero. In the second approach, we established that p had the rather general property of *being negative somewhere*, and used the intermediate value theorem to finish off the argument.

There are many situations like this throughout mathematics, and as they arise certain general properties become established as particularly useful. For example, if you know that a positive integer n is prime, or that a group G is Abelian (that is, $gh = hg$ for any two elements g and h of G), or that a function taking complex numbers to complex numbers is HOLOMORPHIC [I.3 §5.6], then as a consequence of these general properties you know a lot more about the objects in question.

Once properties have established themselves as important, they give rise to a large class of mathematical questions of the following form: given a mathematical structure and a selection of interesting properties that it might have, which combinations of these properties imply which other ones? Not all such questions are interesting, of course—many of them turn out to be quite easy and others are too artificial—but some of them are very natural and surprisingly resistant to one's initial attempts to solve them. This is usually a sign that one has stumbled on what mathematicians would call a “deep” question. In the rest of this section let us look at a problem of this kind.

A group G is called *finitely generated* if there is some finite set $\{x_1, x_2, \dots, x_k\}$ of elements of G such that all the rest can be written as products of elements in that set. For example, the group $SL_2(\mathbb{Z})$ consists of all 2×2 matrices $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ such that a, b, c , and d are integers and $ad - bc = 1$. This group is finitely generated: it is a nice exercise to show that every such matrix can be built from the four matrices $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$, $\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$, and $\begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}$ using matrix multiplication. (See [I.3 §4.2] for a discussion of matrices. A first step toward proving this result is to show that $\begin{pmatrix} 1 & m \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & m+n \\ 0 & 1 \end{pmatrix}$.)

Now let us consider a second property. If x is an element of a group G , then x is said to have *finite order* if there is some power of x that equals the identity. The smallest such power is called the *order of x* . For example, in the multiplicative group of integers mod 7, the identity is 1, and the order of the element 4 is 3, because $4^1 = 4$, $4^2 = 16 \equiv 2$ and $4^3 = 64 \equiv 1 \pmod{7}$. As for 3, its first six powers are 3, 2, 6, 4, 5, 1, so it has order 6. Now some groups have the very special property that there is some integer n such that x^n equals the identity for every x —or, equivalently, the order of every x is a factor of n . What can we say about such groups?

Let us look first at the case where all elements have order 2. Writing e for the identity element, we are assuming that $a^2 = e$ for every element a . If we multiply both sides of this equation by the inverse a^{-1} , then we deduce that $a = a^{-1}$. The opposite implication is equally easy, so such groups are ones where every element is its own inverse.

Now let a and b be two elements of G . For any two elements a and b of any group we have the identity $(ab)^{-1} = b^{-1}a^{-1}$ (simply because $abb^{-1}a^{-1} = aa^{-1} = e$), and in our special group where all elements equal their inverses we can deduce from this that $ab = ba$. That is, G is automatically Abelian.

Already we have shown that one general property, that every element of G squares to the identity, implies another, that G is Abelian. Now let us add the condition that G is finitely generated, and let x_1, x_2, \dots, x_k be a *minimal* set of generators. That is, suppose that every element of G can be built up out of the x_i and that we need all of the x_i to be able to do this. Because G is Abelian and because every element is equal to its own inverse, we can rearrange products of the x_i into a *standard form*, where each x_i occurs at most once and the indices increase. For example, take the product $x_4x_3x_1x_4x_4x_1x_3x_1x_5$. Because G is Abelian, this equals $x_1x_1x_1x_3x_3x_4x_4x_4x_5$, and because each element is its own inverse this equals $x_1x_4x_5$, the standard form of the original expression.

This shows that G can have at most 2^k elements, since for each x_i we have the choice of whether or not to include it in the product (after it has been put in the form above). In particular, the properties “ G is finitely generated” and “every nonidentity element of G has order 2” imply the third property “ G is finite.” It turns out to be fairly easy to prove that two elements whose standard forms are different are themselves different, so in fact G has exactly 2^k elements (where k is the size of a minimal set of generators).

Now let us ask what happens if n is some integer greater than 2 and $x^n = e$ for every element x . That is, if G is finitely generated and $x^n = e$ for every x , must G be finite? This turns out to be a much harder question, originally asked by BURNSIDE [VI.59]. Burnside himself showed that G must be finite if $n = 3$, but it was not until 1968 that his problem was solved, when Adian and Novikov proved the remarkable result that if $n \geq 4381$ then G does *not* have to be finite. There is of course a big gap between 3 and 4381, and progress in bridging it has been slow. It was only in 1992 that this was improved to $n \geq 13$, by Ivanov. And to give an idea of how hard the Burnside problem is, it is still not known whether a

group with two generators such that the fifth power of every element is the identity must be finite.

8 Working with Arguments that Are Not Fully Rigorous

A mathematical statement is considered to be established when it has a proof that meets the high standards of rigor that are characteristic of the subject. However, nonrigorous arguments have an important place in mathematics as well. For example, if one wishes to apply a mathematical statement to another field, such as physics or engineering, then the truth of the statement is often more important than whether one has proved it.

However, this raises an obvious question: if one has not proved a statement, then what grounds could there be for believing it? There are in fact several different kinds of nonrigorous justification, so let us look at some of them.

8.1 Conditional Results

As was mentioned earlier in this article, the Riemann hypothesis is the most famous unsolved problem in mathematics. Why is it considered so important? Why, for example, is it considered more important than the twin-prime conjecture, another problem to do with the behavior of the sequence of primes?

The main reason, though not the only one, is that it and its generalizations have a huge number of interesting consequences. In broad terms, the Riemann hypothesis tells us that the appearance of a certain degree of “randomness” in the sequence of primes is not misleading: in many respects, the primes really do behave like an appropriately chosen random set of integers.

If the primes behave in a random way, then one might imagine that they would be hard to analyze, but in fact randomness can be an advantage. For example, it is randomness that allows me to be confident that at least one girl was born in London on every day of the twentieth century. If the sex of babies were less random, I would be less sure: there could be some strange pattern such as girls being born on Mondays to Thursdays and boys on Fridays to Sundays. Similarly, if I know that the primes behave like a random sequence, then I know a great deal about their average behavior in the long term. The Riemann hypothesis and its generalizations formulate in a precise way the idea that the primes, and other important sequences that arise in number theory, “behave randomly.” That is why they have so many consequences. There are large numbers of papers with theorems that are proved only under the assumption of some version

of the Riemann hypothesis. Therefore, anybody who proves the Riemann hypothesis will change the status of all these theorems from conditional to fully proved.

How should one regard a proof if it relies on the Riemann hypothesis? One could simply say that the proof establishes that such and such a result is implied by the Riemann hypothesis and leave it at that. But most mathematicians take a different attitude. They believe the Riemann hypothesis, and believe that it will one day be proved. So they believe all its consequences as well, even if they feel more secure about results that can be proved unconditionally.

Another example of a statement that is generally believed and used as a foundation for a great deal of further research comes from theoretical computer science. As was mentioned in section 6.4(iv), one of the main aims of computer science is to establish how quickly certain tasks can be performed by a computer. This aim splits into two parts: finding algorithms that work in as few steps as possible, and proving that every algorithm must take at least some particular number of steps. The second of these tasks is notoriously difficult: the best results known are far weaker than what is believed to be true.

There is, however, a class of computational problems, called *NP-complete* problems, that are known to be of *equivalent* difficulty. That is, an efficient algorithm for one of these problems can be converted into an efficient algorithm for any other. Furthermore, it is almost universally believed that there is in fact no efficient algorithm for any of the problems, or, as it is usually expressed, that “P does not equal NP.” Therefore, if you want to demonstrate that no quick algorithm exists for some problem, all you have to do is prove that it is at least as hard as some problem that is already known to be NP-complete. This will not be a rigorous proof, but it will be a convincing demonstration, since most mathematicians are convinced that P does not equal NP. (See COMPUTATIONAL COMPLEXITY [IV.21] for much more on this topic.)

Some areas of research depend on several conjectures rather than just one. It is as though researchers in such areas have discovered a beautiful mathematical landscape and are impatient to map it out despite the fact that there is a great deal that they do not understand. And this is often a very good research strategy, even from the perspective of finding rigorous proofs. There is far more to a conjecture than simply a wild guess: for it to be accepted as important, it should have been subjected to tests of many kinds. For example, does it have consequences that are already known to be true?

Are there special cases that one can prove? If it were true, would it help one solve other problems? Is it supported by numerical evidence? Does it make a bold, precise statement that would probably be easy to refute if it were false? It requires great insight and hard work to produce a conjecture that passes all these tests, but if one succeeds, one has not just an isolated statement, but a statement with numerous connections to other statements. This increases the chances that it will be proved, and greatly increases the chances that the proof of one statement will lead to proofs of others as well. Even a *counterexample* to a good conjecture can be extraordinarily revealing: if the conjecture is related to many other statements, then the effects of the counterexample will permeate the whole area.

One area that is full of conjectural statements is ALGEBRAIC NUMBER THEORY [IV.3]. In particular, the Langlands program is a collection of conjectures, due to Robert Langlands, that relate number theory to representation theory (it is discussed in REPRESENTATION THEORY [IV.12 §6]). Between them, these conjectures generalize, unify, and explain large numbers of other conjectures and results. For example, the Shimura-Taniyama-Weil conjecture, which was central to Andrew Wiles's proof of FERMAT'S LAST THEOREM [V.12], forms one small part of the Langlands program. The Langlands program passes the tests for a good conjecture supremely well, and has for many years guided the research of a large number of mathematicians.

Another area of a similar nature is known as MIRROR SYMMETRY [IV.14]. This is a sort of DUALITY [III.19] that relates objects known as CALABI-YAU MANIFOLDS [III.6], which arise in ALGEBRAIC GEOMETRY [IV.7] and also in STRING THEORY [IV.13 §2], to other, dual manifolds. Just as certain differential equations can become much easier to solve if one looks at the FOURIER TRANSFORMS [III.27] of the functions in question, so there are calculations arising in string theory that look impossible until one transforms them into equivalent calculations in the dual, or "mirror," situation. There is at present no rigorous justification for the transformation, but this process has led to complicated formulas that nobody could possibly have guessed, and some of these formulas have been rigorously proved in other ways. Maxim Kontsevich has proposed a precise conjecture that would explain the apparent successes of mirror symmetry.

8.2 Numerical Evidence

The GOLDBACH CONJECTURE [V.30] states that every even number greater than or equal to 4 is the sum of two

primes. It seems to be well beyond what anybody could hope to prove with today's mathematical machinery, even if one is prepared to accept statements such as the Riemann hypothesis. And yet it is regarded as almost certainly true.

There are two principal reasons for believing Goldbach's conjecture. The first is a reason we have already met: one would expect it to be true if the primes are "randomly distributed." This is because if n is a large even number, then there are many ways of writing $n = a + b$, and there are enough primes for one to expect that from time to time both a and b would be prime.

Such an argument leaves open the possibility that for some value of n that is not too large one might be unlucky, and it might just happen that $n - a$ was composite whenever a was prime. This is where numerical evidence comes in. It has now been checked that every even number up to 10^{14} can be written as a sum of two primes, and once n is greater than this, it becomes extremely unlikely that it could "just happen," by a fluke, to be a counterexample.

This is perhaps rather a crude argument, but there is a way to make it even more convincing. If one makes more precise the idea that the primes appear to be randomly distributed, one can formulate a stronger version of Goldbach's conjecture that says not only that every even number can be written as a sum of two primes, but also roughly how many ways there are of doing this. For instance, if a and $n - a$ are both prime, then neither is a multiple of 3 (unless they are equal to 3 itself). If n is a multiple of 3, then this merely says that a is not a multiple of 3, but if n is of the form $3m + 1$ then a cannot be of the form $3k + 1$ either (or $n - a$ would be a multiple of 3). So, in a certain sense, it is twice as easy for n to be a sum of two primes if it is a multiple of 3. Taking this kind of information into account, one can estimate in how many ways it "ought" to be possible to write n as a sum of two primes. It turns out that, for every even n , there should be many such representations. Moreover, one's predictions of *how* many are closely matched by the numerical evidence: that is, they are true for values of n that are small enough to be checked on a computer. This makes the numerical evidence much more convincing, since it is evidence not just for Goldbach's conjecture itself, but also for the more general principles that led us to believe it.

This illustrates a general phenomenon: the more precise the predictions that follow from a conjecture, the more impressive it is when they are confirmed by later numerical evidence. Of course, this is true not just of mathematics but of science more generally.

8.3 “Illegal” Calculations

In section 6.3 it was stated that “almost nothing is known” about the average end-to-end distance of an n -step self-avoiding walk. That is a statement with which theoretical physicists would strongly disagree. Instead, they would tell you that the end-to-end distance of a typical n -step self-avoiding walk is somewhere in the region of $n^{3/4}$. This apparent disagreement is explained by the fact that, although almost nothing has been rigorously proved, physicists have a collection of nonrigorous methods that, if used carefully, seem to give correct results. With their methods, they have in some areas managed to establish statements that go well beyond what mathematicians can prove. Such results are fascinating to mathematicians, partly because if one regards the results of physicists as mathematical conjectures then many of them are excellent conjectures, by the standards explained earlier: they are deep, completely unguessable in advance, widely believed to be true, backed up by numerical evidence, and so on. Another reason for their fascination is that the effort to provide them with a rigorous underpinning often leads to significant advances in pure mathematics.

To give an idea of what the nonrigorous calculations of physicists can be like, here is a rough description of a famous argument of Pierre-Gilles de Gennes, which lies behind some of the results (or predictions, if you prefer to call them that) of physicists. In statistical physics there is a model known as the n -vector model, closely related to the Ising and Potts models described in PROBABILISTIC MODELS OF CRITICAL PHENOMENA [IV.26]. At each point of \mathbb{Z}^d one places a unit vector in \mathbb{R}^n . This gives rise to a random configuration of unit vectors, with which one associates an “energy” that increases as the angles between neighboring vectors increase. De Gennes found a way of transforming the self-avoiding walk problem so that it could be regarded as a question about the n -vector model in the case $n = 0$. The 0-vector problem itself does not make obvious sense, since there is no such thing as a unit vector in \mathbb{R}^0 , but de Gennes was nevertheless able to take parameters associated with the n -vector model and show that if you let n converge to zero then you obtained parameters associated with self-avoiding walks. He proceeded to choose other parameters in the n -vector model to derive information about self-avoiding walks, such as the expected end-to-end distance.

To a pure mathematician, there is something very worrying about this approach. The formulas that arise in the n -vector model do not make sense when $n = 0$, so

instead one has to regard them as limiting values when n tends to zero. But n is very clearly a positive integer in the n -vector model, so how can one say that it tends to zero? Is there some way of defining an n -vector model for more general n ? Perhaps, but nobody has found one. And yet de Gennes’s argument, like many other arguments of a similar kind, leads to remarkably precise predictions that agree with numerical evidence. There must be a good reason for this, even if we do not understand what it is.

The examples in this section are just a few illustrations of how mathematics is enriched by nonrigorous arguments. Such arguments allow one to penetrate much further into the mathematical unknown, opening up whole areas of research into phenomena that would otherwise have gone unnoticed. Given this, one might wonder whether rigor is important: if the results established by nonrigorous arguments are clearly true, then is that not good enough? As it happens, there are examples of statements that were “established” by nonrigorous methods and later shown to be false, but the most important reason for caring about rigor is that the understanding one gains from a rigorous proof is frequently deeper than the understanding provided by a nonrigorous one. The best way to describe the situation is perhaps to say that the two styles of argument have profoundly benefited each other and will undoubtedly continue to do so.

9 Finding Explicit Proofs and Algorithms

There is no doubt that the equation $x^5 - x - 13 = 0$ has a solution. After all, if we set $f(x) = x^5 - x - 13$, then $f(1) = -13$ and $f(2) = 17$, so somewhere between 1 and 2 there will be an x for which $f(x) = 0$.

That is an example of a *pure existence argument*—in other words, an argument that establishes that something exists (in this case, a solution to a certain equation), without telling us how to find it. If the equation had been $x^2 - x - 13 = 0$, then we could have used an argument of a very different sort: the formula for quadratic equations tells us that there are precisely two solutions, and it even tells us what they are (they are $(1 + \sqrt{53})/2$ and $(1 - \sqrt{53})/2$). However, there is no similar formula for quintic equations (see THE INSOLUBILITY OF THE QUINTIC [V.24]).

These two arguments illustrate a fundamental dichotomy in mathematics. If you are proving that a mathematical object exists, then sometimes you can do so *explicitly*, by actually describing that object, and sometimes you can do so only *indirectly*, by showing that its nonexistence would lead to a contradiction.

There is also a spectrum of possibilities in between. As it was presented, the argument above showed merely that the equation $x^5 - x - 13$ has a solution between 1 and 2, but it also suggests a method for calculating that solution to any desired accuracy. If, for example, you want to know it to two decimal places, then run through the numbers 1, 1.01, 1.02, ..., 1.99, 2 evaluating f at each one. You will find that $f(1.96)$ is approximately -0.202 and $f(1.97)$ is approximately 0.0914 , so there must be a solution between the two (which the calculations suggest will be closer to 1.97 than to 1.96). And in fact there are much better ways, such as NEWTON'S METHOD [III.4 §2.3], of approximating solutions. For many purposes, a pretty formula for a solution is less important than a method of calculating or approximating it. (See NUMERICAL ANALYSIS [IV.20 §1] for a further discussion of this point.) And if one has a method, its usefulness depends very much on whether it works quickly.

Thus, at one end of the spectrum one has simple formulas that define mathematical objects and can easily be used to find them, at the other one has proofs that establish existence but give no further information, and in between one has proofs that yield algorithms for finding the objects, algorithms that are significantly more useful if they run quickly.

Just as, all else being equal, a rigorous argument is preferable to a nonrigorous one, so an explicit or algorithmic argument is worth looking for even if an indirect one is already established, and for similar reasons: the effort to find an explicit argument very often leads to new mathematical insights. (Less obviously, as we shall soon see, finding *indirect* arguments can also lead to new insights.)

One of the most famous examples of a pure existence argument concerns TRANSCENDENTAL NUMBERS [III.43], which are real numbers that are not roots of any polynomial with integer coefficients. The first person to prove that such numbers existed was LIOUVILLE [VI.38], in 1844. He proved that a certain condition was sufficient to guarantee that a number was transcendental and demonstrated that it is easy to construct numbers satisfying his condition (see LIOUVILLE'S THEOREM AND ROTH'S THEOREM [V.25]). After that, various important numbers such as e and π were proved to be transcendental, but these proofs were difficult. Even now there are many numbers that are almost certainly transcendental but which have not been proved to be transcendental. (See IRRATIONAL AND TRANSCENDENTAL NUMBERS [III.43] for more information about this.)

All the proofs mentioned above were direct and explicit. Then in 1873 CANTOR [VI.53] provided a completely different proof of the existence of transcendental numbers, using his theory of COUNTABILITY [III.11]. He proved that the algebraic numbers were countable and the real numbers uncountable. Since countable sets are far smaller than uncountable sets, this showed that almost every real number (though not necessarily almost every real number you will actually meet) is transcendental.

In this instance, each of the two arguments tells us something that the other does not. Cantor's proof shows that there are transcendental numbers, but it does not provide us with a single example. (Strictly speaking, this is not true: one could specify a way of listing the algebraic numbers and then apply Cantor's famous diagonal argument to that particular list. However, the resulting number would be virtually devoid of meaning.) Liouville's proof is much better in that way, as it gives us a method of constructing several transcendental numbers with fairly straightforward definitions. However, if one knew only the explicit arguments such as Liouville's and the proofs that e and π are transcendental, then one might have the impression that transcendental numbers are numbers of a very special kind. The insight that is completely missing from these arguments, but present in Cantor's proof, is that a *typical* real number is transcendental.

For much of the twentieth century, highly abstract and indirect proofs were fashionable, but in more recent years, especially with the advent of the computer, attitudes have changed. (Of course, this is a very general statement about the entire mathematical community rather than about any single mathematician.) Nowadays, more attention is often paid to the question of whether a proof is explicit, and, if so, whether it leads to an efficient algorithm.

Needless to say, algorithms are interesting in themselves, and not just for the light they shed on mathematical proofs. Let us conclude this section with a brief description of a particularly interesting algorithm that has been developed by several authors over the last few years. It gives a way of computing the volume of a high-dimensional convex body.

A shape K is called *convex* if, given any two points x and y in K , the line segment joining x to y lies entirely inside K . For example, a square or a triangle is convex, but a five-pointed star is not. This concept can be generalized straightforwardly to n dimensions, for any n , as can the notions of area and volume.

Now let us suppose that an n -dimensional convex body K is specified for us in the following sense: we have a computer program that runs quickly and tells us, for each point (x_1, \dots, x_n) , whether or not that point belongs to K . How can we estimate the volume of K ? One of the most powerful methods for problems like this is *statistical*: you choose points at random and see whether they belong to K , basing your estimate of the volume of K on the frequency with which they do. For example, if you wanted to estimate π , you could take a circle of radius 1, enclose it in a square of side-length 2, and choose a large number of points randomly from the square. Each point has a probability $\pi/4$ (the ratio of the area π of the circle to the area 4 of the square) of belonging to the circle, so we can estimate π by taking the proportion of points that fall in the circle and multiplying it by 4.

This approach works quite easily for very low dimensions but as soon as n is at all large it runs into a severe difficulty. Suppose for example that we were to try to use the same method for estimating the volume of an n -dimensional sphere. We would enclose that sphere in an n -dimensional cube, choose points at random in the cube, and see how often they belonged to the sphere as well. However, the ratio of the volume of an n -dimensional sphere to that of an n -dimensional cube that contains it is exponentially small, which means that the number of points you have to pick before even one of them lands in the sphere is exponentially large. Therefore, the method becomes hopelessly impractical.

All is not lost, though, because there is a trick for getting around this difficulty. You define a sequence of convex bodies, K_0, K_1, \dots, K_m , each contained in the next, starting with the convex body whose volume you want to know, and ending with the cube, in such a way that the volume of K_i is always at least half that of K_{i+1} . Then for each i you estimate the ratio of the volumes of K_{i-1} and K_i . The product of all these ratios will be the ratio of the volume of K_0 to that of K_m . Since you know the volume of K_m , this tells you the volume of K_0 .

How do you estimate the ratio of the volumes of K_{i-1} and K_i ? You simply choose points at random from K_i and see how many of them belong to K_{i-1} . However, it is just here that the true subtlety of the problem arises: how do you choose points at random from a convex body K_i that you do not know much about? Choosing a random point in the n -dimensional cube is easy, since all you need to do is independently choose n random numbers x_1, \dots, x_n , each between -1 and 1 . But for a general convex body it is not easy at all.

There is a wonderfully clever idea that gets around this problem. It is to design carefully a random walk that starts somewhere inside the convex body and at each step moves to another point, chosen at random from just a few possibilities. The more random steps of this kind that are taken, the less can be said about where the point is, and if the walk is defined properly, it can be shown that after not too many steps, the point reached is almost purely random. However, the proof is not at all easy. (It is discussed further in HIGH-DIMENSIONAL GEOMETRY AND ITS PROBABILISTIC ANALOGUES [IV.24 §6].)

For further discussion of algorithms and their mathematical importance, see COMPUTATIONAL NUMBER THEORY [IV.5], COMPUTATIONAL COMPLEXITY [IV.21], and THE MATHEMATICS OF ALGORITHM DESIGN [VII.5].

10 What Do You Find in a Mathematical Paper?

Mathematical papers have a very distinctive style, one that became established early in the twentieth century. This final section is a description of what mathematicians actually produce when they write.

A typical paper is usually a mixture of formal and informal writing. Ideally (but by no means always), the author writes a readable introduction that tells the reader what to expect from the rest of the paper. And if the paper is divided into sections, as most papers are unless they are quite short, then it is also very helpful to the reader if each section can begin with an informal outline of the arguments to follow. But the main substance of the paper has to be more formal and detailed, so that readers who are prepared to make a sufficient effort can convince themselves that it is correct.

The object of a typical paper is to establish *mathematical statements*. Sometimes this is an end in itself: for example, the justification for the paper may be that it proves a conjecture that has been open for twenty years. Sometimes the mathematical statements are established in the service of a wider aim, such as helping to explain a mathematical phenomenon that is poorly understood. But either way, mathematical statements are the main currency of mathematics.

The most important of these statements are usually called theorems, but one also finds statements called propositions, lemmas, and corollaries. One cannot always draw sharp distinctions between these kinds of statements, but in broad terms this is what the different words mean. A *theorem* is a statement that you regard as intrinsically interesting, a statement that you

might think of isolating from the paper and telling other mathematicians about in a seminar, for instance. The statements that are the main goals of a paper are usually called theorems. A *proposition* is a bit like a theorem, but it tends to be slightly “boring.” It may seem odd to want to prove boring results, but they can be important and useful. What makes them boring is that they do not surprise us in any way. They are statements that we need, that we expect to be true, and that we do not have much difficulty proving.

Here is a quick example of a statement that one might choose to call a proposition. The ASSOCIATIVE LAW FOR A BINARY OPERATION [I.2 §2.4] “ $*$ ” states that $x * (y * z) = (x * y) * z$. One often describes this law informally by saying that “brackets do not matter.” However, while it shows that we can write $x * y * z$ without fear of ambiguity, it does not show quite so obviously that we can write $a * b * c * d * e$, for example. How do we know that, just because the positions of brackets do not matter when you have three objects, they do not matter when you have more than three?

Many mathematics students go happily through university without noticing that this is a problem. It just seems obvious that the associative law shows that brackets do not matter. And they are basically right: although it is not completely obvious, it is certainly not a surprise and turns out to be easy to prove. Since we often need this simple result and could hardly call it a theorem, we might call it a proposition instead. To get a feel for how to prove it, you might wish to show that the associative law implies that

$$(a * ((b * c) * d)) * e = a * (b * ((c * d) * e)).$$

Then you can try to generalize what it is you are doing.

Often, if you are trying to prove a theorem, the proof becomes long and complicated, in which case if you want anybody to read it you need to make the structure of the argument as clear as possible. One of the best ways of doing this is to identify *subgoals*, which take the form of statements intermediate between your initial assumptions and the conclusion you wish to draw from them. These statements are usually called *lemmas*. Suppose, for example, that you are trying to give a very detailed presentation of the standard proof that $\sqrt{2}$ is irrational. One of the facts you will need is that every fraction p/q is equal to a fraction r/s with r and s not both even, and this fact requires a proof. For the sake of clarity, you might well decide to isolate this proof from the main proof and call the fact a lemma. Then you have split your task into two separate tasks: proving the lemma, and proving the main theorem using the lemma. One

can draw a parallel with computer programming: if you are writing a complicated program, it is good practice to divide your main task into subtasks and write separate mini-programs for them, which you can then treat as “black boxes,” to be called upon by other parts of the program whenever they are useful.

Some lemmas are difficult to prove and are useful in many different contexts, so the most important lemmas can be more important than the least important theorems. However, a general rule is that a result will be called a lemma if the main reason for proving it is in order to use it as a stepping stone toward the proofs of other results.

A *corollary* of a mathematical statement is another statement that follows easily from it. Sometimes the main theorem of a paper is followed by several corollaries, which advertise the strength of the theorem. Sometimes the main theorem itself is labeled a corollary, because all the work of the proof goes into proving a different, less punchy statement from which the theorem follows very easily. If this happens, the author may wish to make clear that the corollary is the main result of the paper, and other authors would refer to it as a theorem.

A mathematical statement is established by means of a *proof*. It is a remarkable feature of mathematics that proofs are possible: that, for example, an argument invented by EUCLID [VI.2] over two thousand years ago can still be accepted today and regarded as a completely convincing demonstration. It took until the late nineteenth and early twentieth centuries for this phenomenon to be properly understood, when the language of mathematics was *formalized* (see THE LANGUAGE AND GRAMMAR OF MATHEMATICS [I.2], and especially section 4, for an idea of what this means). Then it became possible to make precise the notion of a proof as well. From a logician’s point of view a proof is a sequence of mathematical statements, each written in a formal language, with the following properties: the first few statements are the initial assumptions, or *premises*; each remaining statement in the sequence follows from earlier ones by means of logical rules that are so simple that the deductions are clearly valid (for instance rules such as “if $P \wedge Q$ is true then P is true,” where “ \wedge ” is the logical symbol for “and”); and the final statement in the sequence is the statement that is to be proved.

The above idea of a proof is a considerable idealization of what actually appears in a normal mathematical paper under the heading “Proof.” That is because a purely formal proof would be very long and almost impossible

to read. And yet, the fact that arguments can in principle be formalized provides a very valuable underpinning for the edifice of mathematics, because it gives a way of resolving disputes. If a mathematician produces an argument that is strangely unconvincing, then the best way to see whether it is correct is to ask him or her to explain it more formally and in greater detail. This will usually either expose a mistake or make it clearer why the argument works.

Another very important component of mathematical papers is *definitions*. This book is full of them: see in particular part III. Some definitions are given simply because they enable one to speak more concisely. For example, if I am proving a result about triangles and I keep needing to consider the distances between the vertices and the opposite sides, then it is a nuisance to have to say “the distances from A, B, and C to the lines BC, AC, and AB, respectively,” so instead I will probably choose a word like “altitude” and write, “Given a vertex of a triangle, define its *altitude* to be the distance from that vertex to the opposite side.” If I am looking at triangles with obtuse angles, then I will have to be more careful: “Given a vertex A of a triangle ABC, define its *altitude* to be the distance from A to the unique line that passes through B and C.” From then on, I can use the word “altitude” and the exposition of my proof will be much more crisp.

Definitions like this are mere definitions of convenience. When the need arises, it is pretty obvious what to do and one does it. But the really interesting definitions are ones that are far from obvious and that make you think in new ways once you know them. A very good example is the definition of the derivative of a function. If you do not know this definition, you will have no idea how to find out for which nonnegative x the function $f(x) = 2x^3 - 3x^2 - 6x + 1$ takes its smallest value. If you do know it, then the problem becomes a simple exercise. That is perhaps an exaggeration, since you also need to know that the minimum will occur either at 0 or at a point where the derivative vanishes, and you will need to know how to differentiate $f(x)$, but these are simple facts—propositions rather than theorems—and the real breakthrough is the concept itself.

There are many other examples of definitions like this, but interestingly they are more common in some branches of mathematics than in others. Some mathematicians will tell you that the main aim of their research is to find the right definition, after which their whole area will be illuminated. Yes, they will have to write proofs, but if the definition is the one they are looking for, then these proofs will be fairly straightforward.

And yes, there will be problems they can solve with the help of the new definition, but, like the minimization problem above, these will not be central to the theory. Rather, they will demonstrate the power of the definition. For other mathematicians, the main purpose of definitions is to prove theorems, but even very theorem-oriented mathematicians will from time to time find that a good definition can have a major effect on their problem-solving prowess.

This brings us to mathematical problems. The main aim of an article in mathematics is usually to prove theorems, but one of the reasons for *reading* an article is to advance one’s own research. It is therefore very welcome if a theorem is proved by a technique that can be used in other contexts. It is also very welcome if an article contains some good unsolved problems. By way of illustration, let us look at a problem that most mathematicians would not take all that seriously, and try to see what it lacks.

A number is called *palindromic* if its representation in base 10 is a palindrome: some simple examples are 22, 131, and 548 845. Of these, 131 is interesting because it is also a prime. Let us try to find some more prime palindromic numbers. Single-digit primes are of course palindromic, and two-digit palindromic numbers are multiples of 11, so only 11 itself is also a prime. So let us move quickly on to three-digit numbers. Here there turn out to be several examples: 101, 131, 151, 191, 313, 353, 373, 383, 727, 757, 787, 797, 919, and 929. It is not hard to show that every palindromic number with an even number of digits is a multiple of 11, but the palindromic primes do not stop at 929—for example, 10 301 is the next smallest.

And now anybody with a modicum of mathematical curiosity will ask the question: are there infinitely many palindromic primes? This, it turns out, is an unsolved problem. It is believed (on the combined grounds that the primes should be sufficiently random and that palindromic numbers with an odd number of digits do not seem to have any particular reason to be factorizable) that there are, but nobody knows how to prove it.

This problem has the great virtue of being easy to understand, which makes it appealing in the way that FERMAT’S LAST THEOREM [V.12] and GOLDBACH’S CONJECTURE [V.30] are appealing. And yet, it is not a central problem in the way that those two are: most mathematicians would put it into a mental box marked “recreational” and forget about it.

What explains this dismissive attitude? Are the primes not central objects of study in mathematics? Well, yes they are, but palindromic numbers are not. And the

main reason they are not is that the definition of “palindromic” is extremely unnatural. If you know that a number is palindromic, what you know is less a feature of the number itself and more a feature of the particular way that, for accidental historical reasons, we choose to represent it. In particular, the property depends on our choice of the number 10 as our base. For example, if we write 131 in base 3, then it becomes 11212, which is no longer the same when written backwards. By contrast, a prime number is prime however you write it.

This is not quite a complete explanation, since there could conceivably be interesting properties that involved the number 10, or at least some artificial choice of number, in an essential way. For example, the problem of whether there are infinitely many primes of the form $2^n - 1$ is considered interesting, despite the use of the particular number 2. However, the choice of 2 can be justified here: $a^n - 1$ has a factor $a - 1$, so for any larger integer the answer would be no. Moreover, numbers of the form $2^n - 1$ have special properties that make them more likely to be prime. (See COMPUTATIONAL NUMBER THEORY [IV.5] for an explanation of this point.)

But even if we replace 10 by the “more natural” number 2 and look at numbers that are palindromic when written in binary, we still do not obtain a property that would be considered a serious topic for research. Suppose that, given an integer n , we define $r(n)$ to be the reverse of n —that is, the number obtained if you write n in binary and then reverse its digits. Then a palindromic number, in the binary sense, is a number n such that $n = r(n)$. But the function $r(n)$ is very strange and “unmathematical.” For instance, the reverses of the numbers from 1 to 20 are 1, 1, 3, 1, 5, 3, 7, 1, 9, 5, 13, 3, 11, 7, 15, 1, 17, 9, 25, and 5, which gives us a sequence with no obvious pattern. Indeed, when one calculates this sequence, one realizes that it is even more artificial than it at first seemed. One might imagine that the reverse of the reverse of a number is the number itself, but that is not so. If you take the number 10, for example, it is 1010 in binary, so its reverse is 0101, which is the number 5. But this we would normally write as 101, so the reverse of 5 is not 10 but 5. But we cannot solve this problem by deciding to write 5 as 0101, since then we would have the problem that 5 was no longer palindromic, when it clearly ought to be.

Does this mean that nobody would be interested in a proof that there were infinitely many palindromic primes? Not at all. It can be shown quite easily that the number of palindromic numbers less than n is in the region of \sqrt{n} , which is a very small fraction indeed. It is notoriously hard to prove results about primes in sparse

sets like this, so a solution to this conjecture would be a big breakthrough. However, the definition of “palindromic” is so artificial that there seems to be no way of using it in a detailed way in a mathematical proof. The only realistic hope of solving this problem would be to prove a much more general result, of which this would be just one of many consequences. Such a result would be wonderful, and undeniably interesting, but you will not discover it by thinking about palindromic numbers. Instead, you would be better off either trying to formulate a more general question, or else looking at a more natural problem of a similar kind. An example of the latter is this: are there infinitely many primes of the form $m^2 + 1$ for some positive integer m ?

Perhaps the most important feature of a good problem is generality: the solution to a good problem should usually have ramifications beyond the problem itself. A more accurate word for this desirable quality is “generalizability,” since some excellent problems may look rather specific. For example, the statement that $\sqrt{2}$ is irrational looks as though it is about just one number, but once you know how to prove it, you will have no difficulty in proving that $\sqrt{3}$ is irrational as well, and in fact the proof can be generalized to a much wider class of numbers (see ALGEBRAIC NUMBERS [IV.3 §14]). It is quite common for a good problem to look uninteresting until you start to think about it. Then you realize that it has been asked for a reason: it might be the “first difficult case” of a more general problem, or it might be just one well-chosen example of a cluster of problems, all of which appear to run up against the same difficulty.

Sometimes a problem is just a question, but frequently the person who asks a mathematical question has a good idea of what the answer is. A *conjecture* is a mathematical statement that the author firmly believes but cannot prove. As with problems, some conjectures are better than others: as we have already discussed in section 8.1, the very best conjectures can have a major effect on the direction of mathematical research.

T&T note:
must fix the
fact that
folio is not
appearing on
the final
page of
some of the
parts before
CRC!

Part II

The Origins of Modern Mathematics

II.1 From Numbers to Number Systems

Fernando Q. Gouvêa

People have been writing numbers down for as long as they have been writing. In every civilization that has developed a way of recording information, we also find a way of recording numbers. Some scholars even argue that numbers came first.

It is fairly clear that numbers first arose as adjectives: they specified how many or how much of something there was. Thus, it was possible to talk about three apricots, say, long before it was possible to talk about the number 3. But once the concept of “threeness” is on the table, so that the same adjective specifies three fish and three horses, and once a written symbol such as “3” is developed that can be used in all of those instances, the conditions exist for 3 itself to emerge as an independent entity. Once it does, we are doing mathematics.

This process seems to have repeated itself many times when new kinds of numbers have been introduced: first a number is used, then it is represented symbolically, and finally it comes to be conceived as a thing in itself and as part of a system of similar entities.

1 Numbers in Early Mathematics

The earliest mathematical documents we know about go back to the civilizations of the ancient Middle East, in Egypt and in Mesopotamia. In both cultures, a scribal class developed. Scribes were responsible for keeping records, which often required them to do arithmetic and solve simple mathematical problems. Most of the mathematical documents we have from those cultures

seem to have been created for the use of young scribes learning their craft. Many of them are collections of problems, provided with either answers or brief solutions: twenty-five problems about digging trenches in one tablet, twelve problems requiring the solution of a linear equation in another, problems about squares and their sides in a third.

Numbers were used both for counting and for measuring, so a need for fractional numbers must have come up fairly early. Fractions are complicated to write down, and computing with them can be difficult. Hence, the problem of “broken numbers” may well have been the first really challenging mathematical problem. How does one write down fractions? The Egyptians and the Mesopotamians came up with strikingly different answers, both of which are also quite different from the way we write them today.

In Egypt (and later in Greece and much of the Mediterranean world), the fundamental notion was “the n th part,” as in “the third part of six is two.” In this language, one would express the idea of dividing 7 by 3 as, “What is the third part of seven?” The answer is, “Two and the third.” The process was complicated by an additional restriction: one never recorded a final result using more than one of the same kind of part. Thus, the number we would want to express as “two fifth parts” would have to be given as “the third and the fifteenth.”

In Mesopotamia, we find a very different idea, which may have arisen to allow easy conversion between different kinds of units. First of all, the Babylonians had a way to generate symbols for all the numbers from 1 to 59. For larger numbers, they used a positional system much like the one we use today, but based on 60 rather than 10. So something like 1, 20 means one sixty and twenty units, that is, $1 \times 60 + 20 = 80$. The same system was then extended to fractions, so that one half

was represented as thirty sixtieths. It is convenient to mark the beginning of the fractional part with a semicolon, though this and the comma are a modern convention that has no counterpart in the original texts. Then, for example, 1;24,36 means $1 + \frac{24}{60} + \frac{36}{60^2}$, the fraction that we would more usually write as $\frac{141}{100}$, or 1.41. The Mesopotamian way of writing numbers is called a *sexagesimal place-value system* by analogy with the system we use today, which is, of course, a *decimal place-value system*.

Neither of these systems is really equipped to deal well with complicated numbers. In Mesopotamia, for example, only *finite* sexagesimal expressions were employed, so the scribes were not able to write down an exact value for the reciprocal of 7 because there is no finite sexagesimal expression for $\frac{1}{7}$. In practice, this meant that to divide by 7 required finding an approximate answer. The Egyptian “parts” system, on the other hand, can represent any positive rational number, but doing so may require a sequence of denominators that to our eyes looks very complicated. One of the surviving papyri includes problems that look *designed* to produce just such complicated answers. One of these answers is “14, the 4th, the 56th, the 97th, the 194th, the 388th, the 679th, the 776th,” which in modern notation is the fraction $14\frac{28}{97}$. It seems that the joy of computation for its own sake became well-established very early in the development of mathematics.

Mediterranean civilizations preserved both of these systems for a while. Most everyday numbers were specified using the system of “parts.” On the other hand, astronomy and navigation required more precision, so the sexagesimal system was used in those fields. This included measuring time and angles. The fact that we still divide an hour into sixty minutes and a minute into sixty seconds goes back, via the Greek astronomers, to the Babylonian sexagesimal fractions; almost four thousand years later, we are still influenced by the Babylonian scribes.

2 Lengths Are Not Numbers

Things get more complicated with the mathematics of classical Greek and Hellenistic civilizations. The Greeks, of course, are famous for coming up with the first mathematical proofs. They were the first to attempt to do mathematics in a rigorously deductive way, using clear initial assumptions and careful statements. This, perhaps, is what led them to be very careful about numbers and their relations to other magnitudes.

Sometime before the fourth century B.C.E., the Greeks made the fundamental discovery of “incommensurable magnitudes.” That is, they discovered that it is not always possible to express two given lengths as (integer) multiples of a third length. It is not just that lengths and numbers are conceptually distinct things (though this was important too). The Greeks had found a *proof* that one cannot use numbers to represent lengths.

Suppose, they argued, you have two line segments. If their lengths are both given by numbers, then those numbers will at worst involve some fractions. By changing the unit of length, then, we can make sure that both of the lengths correspond to whole numbers. In other words, it must be possible to choose a unit length so that each of our segments consists of a whole number multiple of the unit. The two segments, then, could be “measured together,” i.e., would be “commensurable.”

Now here’s the catch: the Greeks could *prove* that this was not always the case. Their standard example had to do with the side and the diagonal of a square. We do not know exactly how they first established that these two segments are not commensurable, but it might have been something like this: if you subtract the side from the diagonal, you will get a segment shorter than either of them; if both side and diagonal are measured by a common unit, then so is the difference. Now repeat the argument: take the remainder and subtract it from the side until we get a second remainder smaller than the first (it can be subtracted twice, in fact). The second remainder will also be measured by the common unit. It turns out to be quite easy to show that *this process will never terminate; instead, it will produce smaller and smaller remainder segments*. Eventually, the remainder segment will be smaller than the unit that supposedly measures it a whole number of times. That is impossible (no whole number is smaller than 1, after all), and hence we can conclude that the common unit does not, in fact, exist.

Of course, the diagonal does in fact have a length. Today, we would say that if the length of the side is one unit, then the length of the diagonal is $\sqrt{2}$ units, and we would interpret this argument as showing that the number $\sqrt{2}$ is not a fraction. The Greeks did not quite see in what sense $\sqrt{2}$ could be a number. Instead, it was a length, or, even better, the ratio between the length of the diagonal and the length of the side. Similar arguments could be applied to other lengths; for example, they knew that the side of a square of area 1 and a square of area 10 are incommensurable.

T&T note: need to fix clash of fractions here before press.

The conclusion, then, is that lengths are not numbers: instead, they are some other kind of magnitude. But now we are faced with a proliferation of magnitudes: numbers, lengths, areas, angles, volumes, etc. Each of these must be taken as a different kind of quantity, not comparable with the others.

This is a problem for geometry, particularly if we want to measure things. The Greeks solved this problem by relying heavily on the notion of a *ratio*. Two quantities of the same type have a ratio, and this ratio was allowed to be equal to the ratio of two quantities of another type: equality of two ratios was defined using Eudoxus's theory of proportion, the latter being one of the most important and deep ideas of Greek geometry. So, for example, rather than talking about a number called π , which to them would not be a number at all, they would say that "the ratio of the circle to the square on its radius is the same as the ratio of the circumference to the diameter." Notice that one of the two ratios is between two areas, the other between two lengths. The number π itself had no name in Greek mathematics, but the Greeks did compare it with ratios between numbers: ARCHIMEDES [VI.3] showed that it was just a little bit less than the ratio of 22 to 7 and just a little bit more than the ratio of 223 to 71.

Doing things this way seems ungainly to us, but it worked very well. Furthermore, it is philosophically satisfying to conceive of a great variety of magnitudes organized into various kinds (segments, angles, surfaces, etc.). Magnitudes of the same kind can be related to one another by ratios, and ratios can be compared with each other because they are relations perceived by our minds. In fact, the word for ratio, both in Greek and in Latin, is the same as the word for "reason" or "explanation" (*logos* in Greek, *ratio* in Latin). From the beginning, "irrational" (*alogos* in Greek) could mean both "without a ratio" and "unreasonable."

Inevitably, the austere system of the theoretical mathematicians was somewhat disconnected from the everyday needs of people who needed to measure things such as lengths and angles. Astronomers kept right on using sexagesimal approximations, as did mapmakers and other scientists. There was some "leakage" of course: in the first century C.E., Heron of Alexandria wrote a book that reads like an attempt to apply the theoreticians' discoveries to practical measurement. It is to him, for example, that we owe the recommendation to use $\frac{22}{7}$ as an approximation for π . (Presumably, he chose Archimedes' upper bound because it was the simpler number.) In theoretical mathematics, however,

the distinction between numbers and other kinds of magnitudes remained firm.

The history of numbers in the West over the fifteen hundred years that followed the classical Greek period can be seen as having two main themes: first, the Greek compartmentalization between different kinds of quantities was slowly demolished; second, in order to do this the notion of number had to be generalized over and over again.

3 Decimal Place Value

Our system for representing whole numbers goes back, ultimately, to the mathematicians of the Indian subcontinent. Sometime before (probably well before) the fifth century C.E., they created nine symbols to designate the numbers from one to nine and used the position of these symbols to indicate their actual value. So a 3 in the units position meant three, and a 3 in the tens position meant three tens, i.e., thirty. This, of course, is what we still do; the symbols themselves have changed, but not the principle. At about the same time, a place marker was developed to indicate an unoccupied space; this eventually evolved into our zero.

Indian astronomy made extensive use of sines, which are almost never whole numbers. To represent these, a Babylonian-style sexagesimal system was used, with each "sexagesimal unit" being represented using the decimal system. So "thirty-three and a quarter" might be represented as 33 15', i.e., 33 units and 15 "minutes" (sixtieths).

Decimal place-value numeration was passed on from India to the Islamic world fairly early. In the ninth century C.E. in Baghdad, the recently established capital of the caliphate, one finds AL-KHWĀRIZMĪ [VI.5] writing a treatise on numeration in the Indian style, "using nine symbols." Several centuries later, al-Khwārizmī's treatise was translated into Latin. It was so popular and influential in late-medieval Europe that decimal numeration was often referred to as "algorism."

It is worth noting that in al-Khwārizmī's writing zero still had a special status: it was a place holder, not a number. But once we have a symbol, and we start doing arithmetic using these symbols, the distinction quickly disappears. We have to know how to add and multiply numbers by zero in order to multiply multi-digit numbers. In this way, "nothing" slowly became a number.

4 What People Want Is a Number

As Greek culture was displaced by other influences, the practical tradition became more important. One can see this in al-Khwārizmī's other famous book, whose title gave us the word "algebra." The book is actually a compendium of many different kinds of practical or semi-practical mathematics problems. Al-Khwārizmī opens the book with a declaration that tells us at once that we are no longer in the Greek mathematical world: "When I considered what people generally want in calculating, I found that it is always a number."

The first portion of al-Khwārizmī's book deals with quadratic equations and with the algebraic manipulations (done entirely in words, with no symbols whatsoever) needed to deal with them. His procedure is exactly the quadratic formula we still use, which of course requires extracting a square root. But in every example the number whose square root we need to find turns out to be a square, so that the square root is easily found—and al-Khwārizmī does get a number!

At other points in the book, however, we can see that al-Khwārizmī is beginning to think of irrational square roots as number-like entities. He teaches the reader how to manipulate symbols with square roots in them, and gives (in words, of course) examples such as $(20 - \sqrt{200}) + (\sqrt{200} - 10) = 10$. In the second part of the book, which deals with geometry and measurement, one even sees an approximation to a square root: "The product is one thousand eight hundred and seventy-five; take its root, it is the area; it is forty-three and a little."

The mathematicians of medieval Islam were influenced not only by the practical tradition represented by al-Khwārizmī, but also by the Greek tradition, especially EUCLID's [VI.2] *Elements*. One finds in their writing a mixture of Greek precision and a more practical approach to measurement. In Omar Khayyam's *Algebra*, for example, one sees both theorems in the Greek style and the desire for numerical solutions. In his discussion of cubic equations Khayyam manages to find solutions by means of geometric constructions but laments his inability to find numerical values.

Slowly, however, the realm of "number" began to grow. The Greeks might have insisted that $\sqrt{10}$ was not a number, but rather a name for a line segment, the side of a square whose area is 10, or a name for a ratio. Among the medieval mathematicians, both in Islam and in Europe, $\sqrt{10}$ started to behave more and more like a

number, entering into operations and even appearing as the solution of certain problems.

5 Giving Equal Status to All Numbers

The idea of extending the decimal place-value system to include fractions was discovered by several mathematicians. The most influential of these was STEVIN [VI.10], a Flemish mathematician and engineer who popularized the system in a booklet called *De Thiende* ("The tenth"), first published in 1585. By extending place value to tenths, hundredths, and so on, Stevin created the system we still use today. More importantly, he explained how it simplified calculations that involved fractions, and gave many practical applications. The cover page, in fact, announces that the book is for "astrologers, surveyors, measurers of tapestries."

Stevin was certainly aware of some of the issues created by his move. He knew, for example, that the decimal expansion for $\frac{1}{3}$ was infinitely long; his discussion simply says that while it might be more correct to say that the full infinite expansion was the correct representation, in practice it made little difference if we truncated it.

Stevin was also aware that his system provided a way to attach a "number" (meaning a decimal expansion) to every single length. He saw little difference between 1.1764705882 (the beginning of the decimal expansion of $\frac{20}{17}$) and 1.4142135623 (the beginning of the decimal expansion of $\sqrt{2}$). In his *Arithmetic* he boldly declared that all (positive) numbers were squares, cubes, fourth powers, etc., and that roots were just numbers. He also says that "there are no absurd, irrational, irregular, inexplicable, or surd numbers." Those were all terms used for irrational numbers, i.e., numbers that are not fractions.

What Stevin was proposing, then, was to flatten the incredible diversity of "quantities" or "magnitudes" into one expansive notion of number, defined by decimal expansions. He was aware that these numbers could be represented as lengths along a line. This amounted to a fairly clear notion of what we now call the positive real numbers.

Stevin's proposal was made immensely more influential by the invention of logarithms. Like the sine and the cosine, these were practical computational tools. In order to be used, they needed to be tabulated, and the tables were given in decimal form. Very soon, everyone was using decimal representation.

PUP: Tim considered option of 'developed' given by proofreader but strongly prefers 'discovered' as the point was that it was 'discovered' by many mathematicians independently. OK?

It was only much later that it came to be understood what a bold leap this move represented. The positive real numbers are not just a larger number system; they are an *immensely* larger number system, whose internal complexity we still do not fully understand (see SET THEORY [IV.1]).

6 Real, False, Imaginary

Even as Stevin was writing, the next steps were being taken: under the pressure of the theory of equations, negative numbers and complex numbers began to be useful. Stevin himself was already aware of negative numbers, though he was clearly not quite comfortable with them. For example, he explained that the fact that -3 is a root of $x^2 + x - 6$ really means that 3 is a root of the associated polynomial $x^2 - x - 6$, obtained by replacing x by $-x$ everywhere.

This was an easy dodge, but cubic equations created more difficult problems. The work of several Italian mathematicians of the sixteenth century led to a method for solving cubic equations. As a crucial step, this method involved extracting a square root. The problem was that the number whose root was needed sometimes came out negative.

Up until then, it had always turned out that when an algebraic problem led to the extraction of the square root of a negative number, the problem simply had no solution. But the equation $x^3 = 15x + 4$ clearly *did* have a solution—indeed, $x = 4$ is one—it was just that applying the cubic formula required computing $\sqrt{-121}$.

It was BOMBELLI [VI.8], also a mathematician and engineer, who decided to bite the bullet and just see what happened. In his *Algebra*, published in 1572, he went ahead and computed with this “new kind of radical” and showed that he could find the solution of the cubic in this way. This showed that the cubic formula did indeed work in this case; more importantly, it showed that these strange new numbers could be useful.

It took a while for people to become comfortable with these new quantities. About fifty years later, we find both Albert Girard and DESCARTES [VI.11] saying that equations can have three sorts of roots: true (meaning positive), false (negative), and imaginary. It is not completely clear that they understood that these imaginary roots would be what we now call complex numbers; Descartes, at least, sometimes seems to be saying that an equation of degree n must have n roots, and

that the ones that are neither “true” nor “false” must simply be imagined.

Slowly, however, complex numbers began to be used. They came up in the theory of equations, in debates about the logarithms of negative numbers, and in connection to trigonometry. Their connection with the sine and cosine functions (via the exponential) was turned into a powerful tool by EULER [VI.19] in the eighteenth century. By the middle of the eighteenth century, it was well-known that every polynomial had a complete set of roots in the complex numbers. This result became known as THE FUNDAMENTAL THEOREM OF ALGEBRA [V.15]; it was finally proved to everyone’s satisfaction by GAUSS [VI.26]. Thus, the theory of equations did not seem to require any further extension of the notion of number.

7 Number Systems, Old and New

Since complex numbers are clearly different from real numbers, their presence stimulated people to begin classifying numbers into different kinds. Stevin’s egalitarianism had its impact, but it could not quite erase the fact that whole numbers are nicer than decimals, and that fractions are generally easier to grasp than irrational numbers.

In the nineteenth century, all sorts of new ideas created the need for a more careful look at this classification. In number theory, Gauss and KUMMER [VI.40] started looking at subsets of the complex numbers that behaved in a way analogous to the integers, such as the set of all numbers $a + b\sqrt{-1}$ with a and b both integers. In the theory of equations, GALOIS [VI.41] pointed out that in order to do a careful analysis of the solvability of an equation one must start by agreeing on what numbers count as “rational.” So, for example, he pointed out that in ABEL’s [VI.33] theorem on the unsolvability of the quintic, “rational” meant “expressible as a quotient of polynomials in the symbols used as the coefficients of the equation,” and he noted that the set of all such expressions obeyed the usual rules of arithmetic.

In the eighteenth century, Johann Lambert had established that e and π were irrational, and conjectured that in fact they were *transcendental*, that is, that they were not roots of any polynomial equation. Even the existence of transcendental numbers was not known at the time; LIOUVILLE [VI.39] proved that such numbers exist in 1844. Within a few decades, it was proved that both e and π were transcendental, and later in the century CANTOR [VI.54] showed that in fact the vast major-

ity of real numbers were transcendental. Cantor's discovery highlighted, for the first time, that the system Stevin had popularized contained unexpected depths.

Perhaps the most important change in the concept of number, however, came after HAMILTON's [VI.37] discovery, in 1843, of a completely new number system. Hamilton had noticed that coordinatizing the plane using complex numbers (rather than simply using pairs of real numbers) vastly simplified plane geometry. He set out to find a similar way to parametrize three-dimensional space. This turned out to be impossible, but led Hamilton to a *four*-dimensional system, which he called the QUATERNIONS [III.78]. These behaved much like numbers, with one crucial difference: multiplication was not commutative, that is, if q and q' are quaternions, qq' and $q'q$ are usually *not* the same.

The quaternions were the first system of "hypercomplex numbers," and their appearance generated lots of new questions. Were there other such systems? What counts as a number system? If certain "numbers" can fail to satisfy the commutative law, can we make numbers that break other rules?

In the long run, this intellectual ferment led mathematicians to let go of the vague notion of "number" or "quantity" and to hold on, instead, to the more formal notion of an algebraic structure. Each of the number systems, in the end, is simply a set of entities on which we can do operations. What makes them interesting is that we can use them to parametrize, or coordinatize, systems that interest us. The whole numbers (or *integers*, to give them their latinized formal name), for example, formalize the notion of counting, while the real numbers parametrize the line and serve as the basis for geometry.

By the beginning of the twentieth century, there were many well-known number systems. The integers had pride of place, followed by a nested hierarchy consisting of the rational numbers (i.e., the fractions), the real numbers (Stevin's decimals, now carefully formalized), and the complex numbers. Still more general than the complex numbers were the quaternions. But these were by no means the only systems around. Number theorists worked with several different fields of algebraic numbers, subsets of the complex numbers that could be understood as autonomous systems. Galois had introduced finite systems that obeyed the usual rules of arithmetic, which we now call finite fields. Function theorists worked with fields of functions; they certainly did not think of these as numbers, but their analogy to number systems was known and exploited.

Early in the twentieth century, Kurt Hensel introduced the p -adic numbers [III.53], which were built from the rational numbers by giving a special role to a prime number p . (Since p can be chosen at will, Hensel in fact created infinitely many new number systems.) These too "obeyed the usual rules of arithmetic," in the sense that addition and multiplication behaved as expected; in modern language, they were *fields*. The p -adics provided the first system of things that were recognizably numbers but that had no visible relation to the real or complex numbers—apart from the fact that both systems contained the rational numbers. As a result, they led Ernst Steinitz to create an abstract theory of fields.

The move to abstraction that appears in Steinitz's work had also occurred in other parts of mathematics, most notably the theory of groups and their representations and the theory of algebraic numbers. All of these theories were brought together into conceptual unity by NOETHER [VI.76], whose program came to be known as "abstract algebra." This left numbers behind completely, focusing instead on the abstract structure of sets with operations.

Today, it is no longer that easy to decide what counts as a "number." The objects from the original sequence of "integer, rational, real, and complex" are certainly numbers, but so are the p -adics. The quaternions are rarely referred to as "numbers," on the other hand, though they can be used to coordinatize certain mathematical notions. In fact, even stranger systems can show up as coordinates, such as Cayley's OCTONIONS [III.78]. In the end, whatever serves to parametrize or coordinatize the problem at hand is what we use. If the requisite system turns out not to exist yet, well, one just has to invent it.

Further Reading

- Berlinghoff, W. P., and F. Q. Gouvêa. 2004. *Math through the Ages: A Gentle History for Teachers and Others*, expanded edn. Farmington, ME/Washington, DC: Oton House/The Mathematical Association of America.
- Ebbinghaus, H.-D., et al. 1991. *Numbers*. New York: Springer.
- Fauvel, J., and J. J. Gray, eds. 1987. *The History of Mathematics: A Reader*. Basingstoke: Macmillan.
- Fowler, D. 1985. 400 years of decimal fractions. *Mathematics Teaching* 110:20–21.
- . 1999. *The Mathematics of Plato's Academy*, 2nd edn. Oxford: Oxford University Press.
- Gouvêa, F. Q. 2003. *p -adic Numbers: An Introduction*, 2nd edn. New York: Springer.

- Katz, V. J. 1998. *A History of Mathematics*, 2nd edn. Reading, MA: Addison-Wesley.
- , ed. 2007. *The Mathematics of Egypt, Mesopotamia, China, India, and Islam: A Sourcebook*. Princeton, NJ: Princeton University Press.
- Mazur, B. 2002. *Imagining Numbers (Particularly the Square Root of Minus Fifteen)*. New York: Farrar, Straus and Giroux.
- Menninger, K. 1992. *Number Words and Number Symbols: A Cultural History of Numbers*. New York: Dover. (Translated by P. Broneer from the revised German edition of 1957/58: *Zahlwort und Ziffer. Eine Kulturgeschichte der Zahl*. Göttingen: Vandenhoeck und Ruprecht.)
- Reid, C. 2006. *From Zero to Infinity: What Makes Numbers Interesting*. Natick, MA: A. K. Peters.

II.2 Geometry

Jeremy Gray

1 Introduction

The modern view of geometry was inspired by the novel geometrical theories of HILBERT [VI.63] and Einstein in the early years of the twentieth century, which built in their turn on other radical reformulations of geometry in the nineteenth century. For thousands of years, the geometrical knowledge of the Greeks, as set out most notably in EUCLID's [VI.2] *Elements*, was held up as a paradigm of perfect rigor, and indeed of human knowledge. The new theories amounted to the overthrow of an entire way of thinking. This essay will pursue the history of geometry, starting from the time of Euclid, continuing with the advent of non-Euclidean geometry, and ending with the work of RIEMANN [VI.49], KLEIN [VI.57], and POINCARÉ [VI.61]. Along the way, we shall examine how and why the notions of geometry changed so remarkably. Modern geometry itself will be discussed in later parts of this book.

2 Naïve Geometry

Geometry generally, and Euclidean geometry in particular, is informally and rightly taken to be the mathematical description of what you see all around you: a space of three dimensions (left-right, up-down, forwards-backwards) that seems to extend indefinitely far. Objects in it have positions, they sometimes move around and occupy other positions, and all of these positions can be specified by measuring lengths along straight lines: this object is twenty meters from that one, it is two meters tall, and so on. We can also measure angles, and there is a subtle relationship between

angles and lengths. Indeed, there is another aspect to geometry, which we do not see but which we reason about. Geometry is a mathematical subject that is full of *theorems*—the isosceles triangle theorem, the Pythagorean theorem, and so on—which collectively summarize what we can say about lengths, angles, shapes, and positions. What distinguishes this aspect of geometry from most other kinds of science is its highly deductive nature. It really seems that by taking the simplest of concepts and thinking hard about them one can build up an impressive, deductive body of knowledge about space without having to gather experimental evidence.

But can we? Is it really as simple as that? Can we have genuine knowledge of space without ever leaving our armchairs? It turns out that we cannot: there are other geometries, also based on the concepts of length and angle, that have every claim to be useful, but that disagree with Euclidean geometry. This is an astonishing discovery of the early nineteenth century, but, before it could be made, a naïve understanding of fundamental concepts, such as straightness, length, and angle, had to be replaced by more precise definitions—a process that took many hundreds of years. Once this had been done, first one and then infinitely many new geometries were discovered.

3 The Greek Formulation

Geometry can be thought of as a set of useful facts about the world, or else as an organized body of knowledge. Either way, the origins of the subject are much disputed. It is clear that the civilizations of Egypt and Babylonia had at least some knowledge of geometry—otherwise, they could not have built their large cities, elaborate temples, and pyramids. But not only is it difficult to give a rich and detailed account of what was known before the Greeks, it is difficult even to make sense of the few scattered sources that we have from before the time of Plato and Aristotle. One reason for this is the spectacular success of the later Greek writer, and author of what became the definitive text on geometry, Euclid of Alexandria (ca. 300 B.C.E.). One glance at his famous *Elements* shows that a proper account of the history of geometry will have to be about something much more than the acquisition of geometrical facts. The *Elements* is a highly organized, deductive body of knowledge. It is divided into a number of distinct themes, but each theme has a complex theoretical structure. Thus, whatever the origins of geometry might have been, by the time of Euclid it had

become the paradigm of a logical subject, offering a kind of knowledge quite different from, and seemingly higher than, knowledge directly gleaned from ordinary experience.

Rather, therefore, than attempt to elucidate the early history of geometry, this essay will trace the high road of geometry's claim on our attention: the apparent certainty of mathematical knowledge. It is exactly this claim to a superior kind of knowledge that led eventually to the remarkable discovery of *non-Euclidean geometry*: there are geometries other than Euclid's that are every bit as rigorously logical. Even more remarkably, some of these turn out to provide better models of physical space than Euclidean geometry.

The *Elements* opens with four books on the study of plane figures: triangles, quadrilaterals, and circles. The famous theorem of Pythagoras is the forty-seventh proposition of the first book. Then come two books on the theory of ratio and proportion and the theory of similar figures (scale copies), treated with a high degree of sophistication. The next three books are about whole numbers, and are presumably a reworking of much older material that would now be classified as elementary number theory. Here, for example, one finds the famous result that there are infinitely many prime numbers. The next book, the tenth, is by far the longest, and deals with the seemingly specialist topic of lengths of the form $\sqrt{a} \pm \sqrt{b}$ (to write them as we would). The final three books, where the curious lengths studied in Book X play a role, are about three-dimensional geometry. They end with the construction of the five regular solids and a proof that there are no more. The discovery of the fifth and last had been one of the topics that excited Plato. Indeed, the five regular solids are crucial to the cosmology of Plato's late work the *Timaeus*.

Most books of the *Elements* open with a number of definitions, and each has an elaborate deductive structure. For example, to understand the Pythagorean theorem, one is driven back to previous results, and thence to even earlier results, until finally one comes to rest on basic definitions. The whole structure is quite compelling: reading it as an adult turned the philosopher Thomas Hobbes from incredulity to lasting belief in a single sitting. What makes the *Elements* so convincing is the nature of the arguments employed. With some exceptions, mostly in the number-theoretic books, these arguments use the axiomatic method. That is to say, they start with some very simple axioms that are intended to be self-evidently true, and proceed by purely logical means to deduce theorems from them.

For this approach to work, three features must be in place. The first is that *circularity* should be carefully avoided. That is, if you are trying to prove a statement P and you deduce it from an earlier statement, and deduce that from a yet earlier statement, and so on, then at no stage should you reach the statement P again. That would not prove P from the axioms, but merely show that all the statements in your chain were equivalent. Euclid did a remarkable job in this respect.

The second necessary feature is that the rules of inference should be clear and acceptable. Some geometrical statements seem so obvious that one can fail to notice that they need to be proved: ideally, one should use no properties of figures other than those that have been clearly stated in their definitions, but this is a difficult requirement to meet. Euclid's success here was still impressive, but mixed. On the one hand, the *Elements* is a remarkable work, far outstripping any contemporary account of any of the topics it covers, and capable of speaking down the millennia. On the other, it has little gaps that from time to time later commentators would fill. For example, it is neither explicitly assumed nor proved in the *Elements* that two circles will meet if their centers lie outside each other and the sum of their radii is greater than the distance between their centers. However, Euclid is surprisingly clear that there are rules of inference that are of general, if not indeed universal, applicability, and others that apply to mathematics because they rely on the meanings of the terms involved.

The third feature, not entirely separable from the second, is adequate definitions. Euclid offered two, or perhaps three, sorts of definition. Book I opens with seven definitions of objects, such as "point" and "line," that one might think were primitive and beyond definition, and it has recently been suggested that these definitions are later additions. Then come, in Book I and again in many later books, definitions of familiar figures designed to make them amenable to mathematical reasoning: "triangle," "quadrilateral," "circle," and so on. The postulates of Book I form the third class of definition and are rather more problematic.

Book I states five "common notions," which are rules of inference of a very general sort. For example, "If equals be added to equals, the wholes are equals." The book also has five "postulates," which are more narrowly mathematical. For example, the first of these asserts that one may draw a straight line from any point to any point. One of these postulates, the fifth, became notorious: the so-called *parallel postulate*. It says that

"If a straight line falling on two straight lines make the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which are the angles less than two right angles."

Parallel lines, therefore, are straight lines that do not meet. A helpful rephrasing of Euclid's parallel postulate was introduced by the Scottish editor, Robert Simson. It appears in his edition of Euclid's *Elements* from 1806. There he showed that the parallel postulate is equivalent, if one assumes those parts of the *Elements* that do not depend on it, to the following statement: given any line m in a plane, and any point P in that plane that does not lie on the line m , there is exactly one line n in the plane that passes through the point P and does not meet the line m . From this formulation it is clear that the parallel postulate makes two assertions: given a line and a point as described, a parallel line *exists* and it is *unique*.

It is worth noting that Euclid himself was probably well aware that the parallel postulate was awkward. It asserts a property of straight lines that seems to have made Greek mathematicians and philosophers uncomfortable, and this may be why its appearance in the *Elements* is delayed until proposition 29 of Book I. The commentator Proclus (fifth century C.E.), in his extensive discussion of Book I of the *Elements*, observed that the hyperbola and asymptote get closer and closer as they move outwards, but they never meet. If a line and a curve can do this, why not two lines? The matter needs further analysis. Unfortunately, not much of the *Elements* would be left if mathematicians dropped the parallel postulate and retreated to the consequences of the remaining definitions: a significant body of knowledge depends on it. Most notably, the parallel postulate is needed to prove that the angles in a triangle add up to two right angles—a crucial result in establishing many other theorems about angles in figures, including the Pythagorean theorem.

Whatever claims educators may have made about Euclid's *Elements* down the ages, a significant number of experts knew that it was an unsatisfactory compromise: a useful and remarkably rigorous theory could be had, but only at the price of accepting the parallel postulate. But the parallel postulate was difficult to accept on trust: it did not have the same intuitively obvious feel of the other axioms and there was no obvious way of verifying it. The higher one's standards, the more painful this compromise was. What, the experts asked, was to be done?

One Greek discussion must suffice here. In Proclus's view, if the truth of the parallel postulate was not obvious, and yet geometry was bare without it, then the only possibility was that it was true because it was a theorem. And so he gave it a proof. He argued as follows. Let two lines m and n cross a third line k at P and Q , respectively, and make angles with it that add up to two right angles. Now draw a line l that crosses m at P and enters the space between the lines m and n . The distance between l and m as one moves away from the point P continually increases, said Proclus, and therefore line l must eventually cross line n .

Proclus's argument is flawed. The flaw is subtle, and sets us up for what is to come. He was correct that the distance between the lines l and m increases indefinitely. But his argument assumes that the distance between lines m and n does not *also* increase indefinitely, and is instead bounded. Now Proclus knew very well that *if* the parallel postulate is granted, *then* it can be shown that the lines m and n are parallel and that the distance between them is a constant. But until the parallel postulate is proved, nothing prevents one saying that the lines m and n diverge. Proclus's proof does not therefore work unless one can show that lines that do not meet also do not diverge.

Proclus's attempt was not the only one, but it is typical of such arguments, which all have a standard form. They start by detaching the parallel postulate from Euclid's *Elements*, together with all the arguments and theorems that depend on it. Let us call what remains the "core" of the *Elements*. Using this core, an attempt is then made to derive the parallel postulate as a theorem. The correct conclusion to be derived from Proclus's attempt is not that the parallel postulate is a theorem, but rather that, given the core of the *Elements*, the parallel postulate is equivalent to the statement that lines that do not meet also do not diverge. Aganis, a writer of the sixth century C.E. about whom almost nothing is known, assumed, in a later attempt, that parallel lines are everywhere equidistant, and his argument showed only that, given the core, the Euclidean definition of parallel lines is equivalent to defining them to be equidistant.

Notice that one cannot even enter this debate unless one is clear which properties of straight lines belong to them by definition, and which are to be derived as theorems. If one is willing to add to the store of "common-sense" assumptions about geometry as one goes along, the whole careful deductive structure of the *Elements* collapses into a pile of facts.

This deductive character of the *Elements* is clearly something that Euclid regarded as important, but one can also ask what he thought geometry was *about*. Was it meant, for example, as a mathematical description of space? No surviving text tells us what he thought about this question, but it is worth noting that the most celebrated Greek theory of the universe, developed by Aristotle and many later commentators, assumed that space was finite, bounded by the sphere of the fixed stars. The mathematical space of the *Elements* is infinite, and so one has at least to consider the possibility that, for all these writers, mathematical space was not intended as a simple idealization of the physical world.

4 Arab and Islamic Commentators

What we think of today as Greek geometry was the work of a handful of mathematicians, mostly concentrated in a period of less than two centuries. They were eventually succeeded by a somewhat larger number of Arabic and Islamic writers, spread out over a much greater area and a longer time. These writers tend to be remembered as commentators on Greek mathematics and science, and for transmitting them to later Western authors, but they should also be remembered as creative, innovative mathematicians and scientists in their own right. A number of them took up the study of Euclid's *Elements*, and with it the problem of the parallel postulate. They too took the view that it was not a proper postulate, but one that could be proved as a theorem using the core alone.

Among the first to attempt a proof was Thābit ibn Qurra. He was a pagan from near Aleppo who lived and worked in Baghdad, where he died in 901. Here there is room to describe only his first approach. He argued that if two lines m and n are crossed by a third, k , and if they approach each other on one side of the line k , then they diverge indefinitely on the other side of k . He deduced that two lines that make equal alternate angles with a transversal (the marked angles in figure 1) cannot approach each other on one side of a transversal: the symmetry of the situation would imply that they approached on the other side as well, but he had shown that they would have to diverge on the other side. From this he deduced the Euclidean theory of parallels, but his argument was also flawed, since he had not considered the possibility that two lines could *diverge* in both directions.

The distinguished Islamic mathematician and scientist ibn al-Haytham was born in Basra in 965 and died

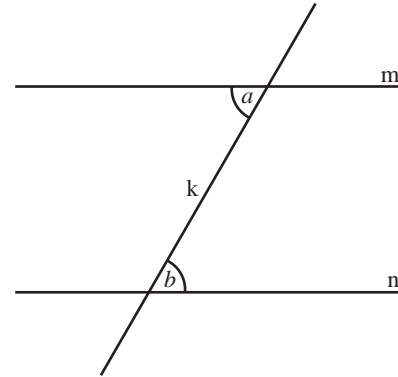


Figure 1 The lines m and n make equal alternate angles a and b with the transversal k .

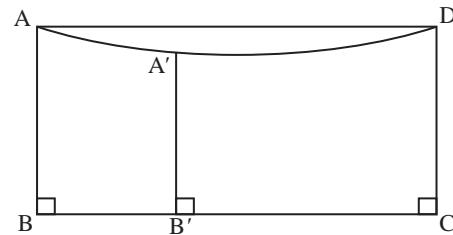


Figure 2 AB and CD are equal, the angle ADC is a right angle, $A'B'$ is an intermediate position of AB as it moves toward CD .

in Egypt in 1041. He took a quadrilateral with two equal sides perpendicular to the base and dropped a perpendicular from one side to the other. He now attempted to prove that this perpendicular is equal to the base, and to do so he argued that as one of two original perpendiculars is moved toward the other, its tip sweeps out a straight line, which will coincide with the perpendicular just dropped (see figure 2). This amounts to the assumption that the curve everywhere equidistant from a straight line is itself straight, from which the parallel postulate easily follows, and so his attempt fails. His proof was later heavily criticized by Omar Khayyam for its use of motion, which he found fundamentally unclear and alien to Euclid's *Elements*. It is indeed quite distinct from any use Euclid had for motion in geometry, because in this case the nature of the curve obtained is not clear: it is precisely what needs to be analyzed.

The last of the Islamic attempts on the parallel postulate is due to Naṣīr al-Dīn al-Ṭūsī. He was born in Iran in 1201 and died in Baghdad in 1274. His extensive

PUP: query relating to figure acted on here and sentence rewritten to clarify things. OK?

commentary is also one of our sources of knowledge of earlier Islamic mathematical work on this subject. Al-Ṭūsī focused on showing that if two lines begin to converge, then they must continue to do so until they eventually meet. To this end he set out to show that

- (*) if l and m are two lines that make an angle of less than a right angle, then every line perpendicular to l meets the line m .

He showed that if (*) is true, then the parallel postulate follows. However, his argument for (*) is flawed.

It is genuinely difficult to see what is wrong with some of these arguments if one uses only the techniques available to mathematicians of the time. Islamic mathematicians showed a degree of sophistication that was not to be surpassed by their Western successors until the eighteenth century. Unfortunately, however, their writings did not come to the attention of the West until much later, with the exception of a single work in the Vatican Library, published in 1594, which was for many years erroneously attributed to al-Ṭūsī (and which may have been the work of his son).

5 The Western Revival of Interest

The Western revival of interest in the parallel postulate came with the second wave of translations of Greek mathematics, led by Commandino and Maurolico in the sixteenth century and spread by the advent of printing. Important texts were discovered in a number of older libraries, and ultimately this led to the production of new texts of Euclid's *Elements*. Many of these had something to say about the problem of parallels, pithily referred to by Henry Savile as "a blot on Euclid." For example, the powerful Jesuit Christopher Clavius, who edited and reworked the *Elements* in 1574, tried to argue that parallel lines could be defined as equidistant lines.

The ready identification of physical space with the space of Euclidean geometry came about gradually during the sixteenth and seventeenth centuries, after the acceptance of Copernican astronomy and the abolition of the so-called sphere of fixed stars. It was canonized by NEWTON [VI.14] in his *Principia Mathematica*, which proposed a theory of gravitation that was firmly situated in Euclidean space. Although Newtonian physics had to fight for its acceptance, Newtonian cosmology had a smooth path and became the unchallenged orthodoxy of the eighteenth century. It can be argued that this identification raised the stakes,

because any unexpected or counterintuitive conclusion drawn solely from the core of the *Elements* was now, possibly, a counterintuitive fact about space.

In 1663 the English mathematician John Wallis took a much more subtle view of the parallel postulate than any of his predecessors. He had been instructed by Halley, who could read Arabic, in the contents of the apocryphal edition of al-Ṭūsī's work in the Vatican Library, and he too gave an attempted proof. Unusually, Wallis also had the insight to see where his own argument was flawed, and commented that what it really showed was that, in the presence of the core, the parallel postulate was equivalent to the assertion that there exist similar figures that are not congruent.

Half a century later, Wallis was followed by the most persistent and thoroughgoing of all the defenders of the parallel postulate, Gerolamo Saccheri, an Italian Jesuit who published in 1733, the year of his death, a short book called *Euclid Freed of Every Flaw*. This little masterpiece of classical reasoning opens with a trichotomy. Unless the parallel postulate is known, the angle sum of a triangle may be either less than, equal to, or greater than two right angles. Saccheri showed that whatever happens in one triangle happens for them all, so there are apparently three geometries compatible with the core. In the first, every triangle has an angle sum less than two right angles (call this case L). In the second, every triangle has an angle sum equal to two right angles (call this case E). In the third, every triangle has an angle sum greater than two right angles (call this case G). Case E is, of course, Euclidean geometry, which Saccheri wished to show was the only case possible. He therefore set to work to show that each of the other cases independently self-destructed. He was successful with case G, and then turned to case L "which alone obstructs the truth of the [parallel] axiom," as he put it.

Case L proved to be difficult, and during the course of his investigations Saccheri established a number of interesting propositions. For example, if case L is true, then two lines that do not meet have just one common perpendicular, and they diverge on either side of it. In the end, Saccheri tried to deal with his difficulties by relying on foolish statements about the behavior of lines at infinity: it was here that his attempted proof failed.

Saccheri's work sank slowly, though not completely, into obscurity. It did, however, come to the attention of the Swiss mathematician Johann Heinrich Lambert, who pursued the trichotomy but, unlike Saccheri,

stopped short of claiming success in proving the parallel postulate. Instead the work was abandoned, and was published only in 1786, after his death. Lambert distinguished carefully between unpalatable results and impossibilities. He had a sketch of an argument to show that in case L the area of a triangle is proportional to the difference between two right angles and the angle sum of the triangle. He knew that in case L similar triangles had to be congruent, which would imply that the tables of trigonometric functions used in astronomy were not in fact valid and that different tables would have to be produced for every size of triangle. In particular, for every angle less than 60° there would be precisely one equilateral triangle with that given angle at each vertex. This would lead to what philosophers called an “absolute” measure of length (one could take, for instance, the length of the side of an equilateral triangle with angles equal to 30°), which LEIBNIZ’S [VI.15] follower Wolff had said was impossible. And indeed it is counterintuitive: lengths are generally defined in relative terms, as, for instance, a certain proportion of the length of a meter rod in Paris, or of the circumference of Earth, or of something similar. But such arguments, said Lambert, “were drawn from love and hate, with which a mathematician can have nothing to do.”

6 The Shift of Focus around 1800

The phase of Western interest in the parallel postulate that began with the publication of modern editions of Euclid’s *Elements* started to decline with a further turn in that enterprise. After the French revolution, LEGENDRE [VI.24] set about writing textbooks, largely for the use of students hoping to enter the École Polytechnique, that would restore the study of elementary geometry to something like the rigorous form in which it appeared in the *Elements*. However, it was one thing to seek to replace books of a heavily intuitive kind, but quite another to deliver the requisite degree of rigor. Legendre, as he came to realize, ultimately failed in his attempt. Specifically, like everyone before him, he was unable to give an adequate defense of the parallel postulate. Legendre’s *Éléments de Géométrie* ran to numerous editions, and from time to time a different attempt on the postulate was made. Some of these attempts would be hard to describe favorably, but the best can be extremely persuasive.

Legendre’s work was classical in spirit, and he still took it for granted that the parallel postulate had to be true. But by around 1800 this attitude was no longer

universally held. Not everybody thought that the postulate must, somehow, be defended, and some were prepared to contemplate with equanimity the idea that it might be false. No clearer illustration of this shift can be found than a brief note sent to GAUSS [VI.26] by F. K. Schweikart, a Professor of Law at the University of Marburg, in 1818. Schweikart described in a page the main results he had been led to in what he called “astral geometry,” in which the angle sum of a triangle was less than two right angles: squares had a particular form, and the altitude of a right-angled isosceles triangle was bounded by an amount Schweikart called “the constant.” Schweikart went so far as to claim that the new geometry might even be the true geometry of space. Gauss replied positively. He accepted the results, and he claimed that he could do all of elementary geometry once a value for the constant was given. One could argue, somewhat ungenerously, that Schweikart had done little more than read Lambert’s posthumous book—although the theorem about isosceles triangles is new. However, what is notable is the attitude of mind: the idea that this new geometry might be true, and not just a mathematical curiosity. Euclid’s *Elements* shackled him no more.

Unfortunately, it is much less clear precisely what Gauss himself thought. Some historians, mindful of Gauss’s remarkable mathematical originality, have been inclined to interpret the evidence in such a way that Gauss emerges as the first person to discover non-Euclidean geometry. The evidence, however, is very slight, and difficult to interpret. There are traces of some early investigations by Gauss of Euclidean geometry that include a study of a new definition of parallel lines; there are claims made by Gauss late in life that he had known this or that fact for many years; and there are letters he wrote to his friends. But there is no material in the surviving papers that allows us to reconstruct what Gauss knew, or that supports the claim that Gauss discovered non-Euclidean geometry.

Rather, the picture would seem to be that Gauss came to realize during the 1810s that all previous attempts to derive the parallel postulate from the core of Euclidean geometry had failed and that all future attempts would probably fail as well. He became more and more convinced that there was another possible geometry of space. Geometry ceased, in his mind, to have the status of arithmetic, which was a matter of logic, and became associated with mechanics, an empirical science. The simplest accurate statement of Gauss’s position through the 1820s is that he did not doubt that

space might be described by a non-Euclidean geometry, and of course there was only one possibility: that of case L described above. It was an empirical matter, but one that could not be resolved by land-based measurements because any departure from Euclidean geometry was, evidently, very small. In this view he was supported by his friends, such as Bessel and Olbers, both professional astronomers. Gauss the scientist was convinced, but Gauss the mathematician may have retained a small degree of doubt, and certainly never developed the mathematical theory required to describe non-Euclidean geometry adequately.

One theory available to Gauss from the early 1820s was that of differential geometry. Gauss eventually published one of his masterworks on this subject, his *Disquisitiones Generales circa Superficies Curvas* (1827). In it he showed how to describe geometry on any surface in space, and how to regard certain features of the geometry of a surface as intrinsic to the surface and independent of how the surface was embedded into three-dimensional space. It would have been possible for Gauss to consider a surface of constant negative CURVATURE [III.80], and to show that triangles on such a surface are described by hyperbolic trigonometric formulas, but he did not do this until the 1840s. Had he done so, he would have had a surface on which the formulas of a geometry satisfying case L apply.

A surface, however, is not enough. We accept the validity of two-dimensional Euclidean geometry because it is a simplification of three-dimensional Euclidean geometry. Before a two-dimensional geometry satisfying the hypotheses of case L can be accepted, it is necessary to show that there is a plausible three-dimensional geometry analogous to case L. Such a geometry has to be described in detail and shown to be as plausible as Euclidean three-dimensional geometry. This Gauss simply never did.

7 Bolyai and Lobachevskii

The fame for discovering non-Euclidean geometry goes to two men, BOLYAI [VI.34] in Hungary and LOBACHEVSKII [VI.31] in Russia, who independently gave very similar accounts of it. In particular, both men described a system of geometry in two and three dimensions that differed from Euclid's but had an equally good claim to be the geometry of space. Lobachevskii published first, in 1829, but only in an obscure Russian journal, and then in French in 1837, in German in 1840, and again in French in 1855. Bolyai published his account in 1831,

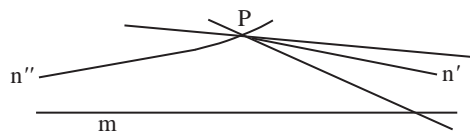


Figure 3 The lines n' and n'' through P separate the lines through P that meet the line m from those that do not.

in an appendix to a two-volume work on geometry by his father.

It is easiest to describe their achievements together. Both men defined parallels in a novel way, as follows. Given a point P and a line m there will be some lines through P that meet m and others that do not. Separating these two sets will be two lines through P that do not quite meet m but which might come arbitrarily close, one to the right of P and one to the left. This situation is illustrated in figure 3: the two lines in question are n' and n'' . Notice that lines on the diagram appear curved. This is because, in order to represent them on a flat, Euclidean page, it is necessary to distort them, unless the geometry is itself Euclidean, in which case one can put n' and n'' together and make a single line that is infinite in both directions.

Given this new way of talking, it still makes sense to talk of dropping the perpendicular from P to the line m . The left and right parallels to m through P make equal angles with the perpendicular, called the *angle of parallelism*. If the angle is a right angle, then the geometry is Euclidean. However, if it is less than a right angle, then the possibility arises of a new geometry. It turns out that the size of the angle depends on the length of the perpendicular from P to m . Neither Bolyai nor Lobachevskii expended any effort in trying to show that there was not some contradiction in taking the angle of parallelism to be less than a right angle. Instead, they simply made the assumption and expended a great deal of effort on determining the angle from the length of the perpendicular.

They both showed that, given a family of lines all parallel (in the same direction) to a given line, and given a point on one of the lines, there is a curve through that point that is perpendicular to each of the lines (figure 4).

In Euclidean geometry the curve defined in this way is the straight line that is at right angles to the family of parallel lines and that passes through the given point (figure 5). If, again in Euclidean geometry, one takes the family of all lines through a common point Q and chooses another point P , then there will be a curve

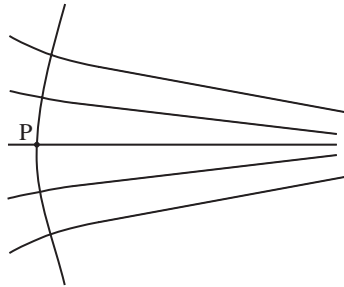


Figure 4 A curve perpendicular to a family of parallels.

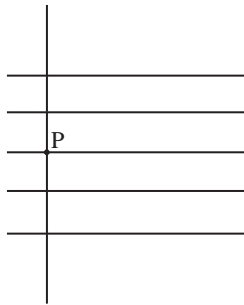


Figure 5 A curve perpendicular to a family of Euclidean parallels.

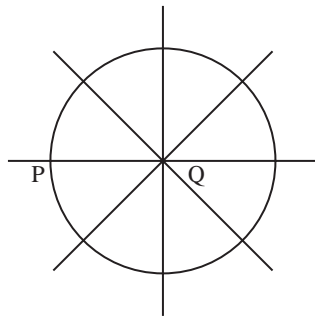


Figure 6 A curve perpendicular to a family of Euclidean lines through a point.

through P that is perpendicular to all the lines: the circle with center Q that passes through P (figure 6).

The curve defined by Bolyai and Lobachevskii has some of the properties of both these Euclidean constructions: it is perpendicular to all the parallels, but it is curved and not straight. Bolyai called such a curve an *L-curve*. Lobachevskii more helpfully called it a *horocycle*, and the name has stuck.

Their complicated arguments took both men into three-dimensional geometry. Here Lobachevskii's argu-

ments were somewhat clearer than Bolyai's, and both men notably surpassed Gauss. If the figure defining a horocycle is rotated about one of the parallel lines, the lines become a family of parallel lines in three dimensions and the horocycle sweeps out a bowl-shaped surface, called the *F-surface* by Bolyai and the horosphere by Lobachevskii. Both men now showed that something remarkable happens. Planes through the horosphere cut it either in circles or in horocycles, and if a triangle is drawn on a horosphere whose sides are horocycles, then the angle sum of such a triangle is two right angles. To put this another way, although the space that contains the horosphere is a three-dimensional version of case L, and is definitely not Euclidean, the geometry you obtain when you restrict attention to the horosphere is (two-dimensional) Euclidean geometry!

Bolyai and Lobachevskii also knew that one can draw spheres in their three-dimensional space, and they showed (though in this they were not original) that the formulas of spherical geometry hold independently of the parallel postulate. Lobachevskii now used an ingenious construction involving his parallel lines to show that a triangle on a sphere determines and is determined by a triangle in the plane, which also determines and is determined by a triangle on the horosphere. This implies that the formulas of spherical geometry must determine formulas that apply to the triangles on the horosphere. On checking through the details, Lobachevskii, and in more or less the same way Bolyai, showed that the triangles on the horosphere are described by the formulas of hyperbolic trigonometry.

The formulas for spherical geometry depend on the radius of the sphere in question. Similarly, the formulas of hyperbolic trigonometry depend on a certain real parameter. However, this parameter does not have a similarly clear geometrical interpretation. That defect apart, the formulas have a number of reassuring properties. In particular, they closely approximate the familiar formulas of plane geometry when the sides of the triangles are very small, which helps to explain how this geometry could have remained undetected for so long—it differs very little from Euclidean geometry in small regions of space. Formulas for length and area can be developed in the new setting: they show that the area of a triangle is proportional to the amount by which the angle sum of the triangle falls short of two right angles. Lobachevskii, in particular, seems to have felt that the very fact that there were neat and plausible formulas of this kind was enough reason to accept the new geometry. In his opinion, all geometry was about

measurement, and theorems in geometry were unfailing connections between measurements expressed by formulas. His methods produced such formulas, and that, for him, was enough.

Bolyai and Lobachevskii, having produced a description of a novel three-dimensional geometry, raised the question of which geometry is true: is it Euclidean geometry or is it the new geometry for some value of the parameter that could presumably be determined experimentally? Bolyai left matters there, but Lobachevskii explicitly showed that measurements of stellar parallax might resolve the question. Here he was unsuccessful: such experiments are notoriously delicate.

By and large, the reaction to Bolyai and Lobachevskii's ideas during their lifetimes was one of neglect and hostility, and they died unaware of the success their discoveries would ultimately have. Bolyai and his father sent their work to Gauss, who replied in 1832 that he could not praise the work "for to do so would be to praise myself," adding, for extra measure, a simpler proof of one of Janos Bolyai's opening results. He was, he said, nonetheless delighted that it was the son of his old friend who had taken precedence over him. Janos Bolyai was enraged, and refused to publish again, thus depriving himself of the opportunity to establish his priority over Gauss by publishing his work as an article in a mathematics journal. Oddly, there is no evidence that Gauss knew the details of the young Hungarian's work in advance. More likely, he saw at once how the theory would go once he appreciated the opening of Bolyai's account.

A charitable interpretation of the surviving evidence would be that, by 1830, Gauss was convinced of the possibility that physical space might be described by non-Euclidean geometry, and he surely knew how to handle two-dimensional non-Euclidean geometry using hyperbolic trigonometry (although no detailed account of this survives from his hand). But the three-dimensional theory was known first to Bolyai and Lobachevskii, and may well not have been known to Gauss until he read their work.

Lobachevskii fared little better than Bolyai. His initial publication of 1829 was savaged in the press by Ostrogradskii, a much more established figure who was, moreover, in St Petersburg, whereas Lobachevskii was in provincial Kazan. His account in *Journal für die reine und angewandte Mathematik* (otherwise known as *Crelle's Journal*) suffered grievously from referring to results proved only in the Russian papers from which it had been adapted. His booklet of 1840 drew

only one review, of more than usual stupidity. He did, however, send it to Gauss, who found it excellent and had Lobachevskii elected to the Göttingen Academy of Sciences. But Gauss's enthusiasm stopped there, and Lobachevskii received no further support from him.

Such a dreadful response to a major discovery invites analysis on several levels. It has to be said that the definition of parallels upon which both men depended was, as it stood, inadequate, but their work was not criticized on that account. It was dismissed with scorn, as if it were self-evident that it was wrong: so wrong that it would be a waste of time finding the error it surely contained, so wrong that the right response was to heap ridicule upon its authors or simply to dismiss them without comment. This is a measure of the hold that Euclidean geometry still had on the minds of most people at the time. Even Copernicanism, for example, and the discoveries of Galileo drew a better reception from the experts.

8 Acceptance of Non-Euclidean Geometry

When Gauss died in 1855, an immense amount of unpublished mathematics was found among his papers. Among it was evidence of his support for Bolyai and Lobachevskii, and his correspondence endorsing the possible validity of non-Euclidean geometry. As this was gradually published, the effect was to send people off to look for what Bolyai and Lobachevskii had written and to read it in a more positive light.

PUP: Tim thinks 'among' is better and clearer than 'in' (suggested by proofreader). OK?

Quite by chance, Gauss had also had a student at Göttingen who was capable of moving the matter decisively forward, even though the actual amount of contact between the two was probably quite slight. This was RIEMANN [VI.49]. In 1854 he was called to defend his Habilitation thesis, the postdoctoral qualification that was a German mathematician's license to teach in a university. As was the custom, he offered three titles and Gauss, who was his examiner, chose the one Riemann least expected: "On the hypotheses that lie at the foundation of geometry." The paper, which was to be published only posthumously, in 1867, was nothing less than a complete reformulation of geometry.

Riemann proposed that geometry was the study of what he called MANIFOLDS [I.3 §§6.9, 6.10]. These were "spaces" of points, together with a notion of distance that looked like Euclidean distance on small scales but which could be quite different at larger scales. This kind of geometry could be done in a variety of ways, he suggested, by means of the calculus. It could be carried

out for manifolds of any dimension, and in fact Riemann was even prepared to contemplate manifolds for which the dimension was infinite.

A vital aspect of Riemann's geometry, in which he followed the lead of Gauss, was that it was concerned only with those properties of the manifold that were *intrinsic*, rather than properties that depended on some embedding into a larger space. In particular, the distance between two points x and y was defined to be the length of the shortest curve joining x and y that lay entirely within the surface. Such curves are called *geodesics*. (On a sphere, for example, the geodesics are arcs of great circles.)

Even two-dimensional manifolds could have different, intrinsic curvatures—indeed, a single two-dimensional manifold could have different curvatures in different places—so Riemann's definition led to infinitely many genuinely distinct geometries in each dimension. Furthermore, these geometries were best defined without reference to a Euclidean space that contained them, so the hegemony of Euclidean geometry was broken once and for all.

As the word “hypotheses” in the title of his thesis suggests, Riemann was not at all interested in the sorts of assumptions needed by Euclid. Nor was he much interested in the opposition between Euclidean and non-Euclidean geometry. He made a small reference at the start of his paper to the murkiness that lay at the heart of geometry, despite the efforts of Legendre, and toward the end he considered the three different geometries on two-dimensional manifolds for which the curvature is constant. He noted that one was spherical geometry, another was Euclidean geometry, and the third was different again, and that in each case the angle sums of all triangles could be calculated as soon as one knew the sum of the angles of any one triangle. But he made no reference to Bolyai or Lobachevskii, merely noting that if the geometry of space was indeed a three-dimensional geometry of constant curvature, then to determine which geometry it was would involve taking measurements in unfeasibly large regions of space. He did discuss generalizations of Gauss's curvature to spaces of arbitrary dimension, and he showed what METRICS [III.58] (that is, definitions of distance) there could be on spaces of constant curvature. The formula he wrote down is very general, but as with Bolyai and Lobachevskii it depended on a certain real parameter—the curvature. When the curvature is negative, his definition of distance gives a description of non-Euclidean geometry.

Riemann died in 1866, and by the time his thesis was published an Italian mathematician, Eugenio Beltrami, had independently come to some of the same ideas. He was interested in what the possibilities were if one wished to map one surface to another. For example, one might ask, for some particular surface S , whether it is possible to find a map from S to the plane such that the geodesics in S are mapped to straight lines in the plane. He found that the answer was yes if and only if the space has constant curvature. There is, for example, a well-known map from the hemisphere to a plane with this property. Beltrami found a simple way of modifying the formula so that now it defined a map from a surface of constant *negative* curvature onto the interior of a disk, and he realized the significance of what he had done: his map defined a metric on the interior of the disk, and the resulting metric space obeyed the axioms for non-Euclidean geometry; therefore, those axioms would not lead to a contradiction.

Some years earlier, Minding, in Germany, had found a surface, sometimes called the pseudosphere, that had constant negative curvature. It was obtained by rotating a curve called the tractrix about its axis. This surface has the shape of a bugle, so it seemed rather less natural than the space of Euclidean plane geometry and unsuitable as a rival to it. The pseudosphere was independently rediscovered by LIOUVILLE [VI.39] some years later, and Codazzi learned of it from that source and showed that triangles on this surface are described by the formulas of hyperbolic trigonometry. But none of these men saw the connection to non-Euclidean geometry—that was left to Beltrami.

Beltrami realized that his disk depicted an infinite space of constant negative curvature, in which the geometry of Lobachevskii (he did not know at that time of Bolyai's work) held true. He saw that it related to the pseudosphere in a way similar to the way that a plane relates to an infinite cylinder. After a period of some doubt, he learned of Riemann's ideas and realized that his disk was in fact as good a depiction of the space of non-Euclidean geometry as any could be; there was no need to realize his geometry as that of a surface in Euclidean three-dimensional space. He thereupon published his essay, in 1868. This was the first time that sound foundations had been publicly given for the area of mathematics that could now be called non-Euclidean geometry.

In 1871 the young KLEIN [VI.57] took up the subject. He already knew that the English mathematician CAYLEY [VI.46] had contrived a way of introducing

Euclidean metrical concepts into PROJECTIVE GEOMETRY [I.3 §6.7]. While studying at Berlin, Klein saw a way of generalizing Cayley's idea and exhibiting Beltrami's non-Euclidean geometry as a special case of projective geometry. His idea met with the disapproval of WEIERSTRASS [VI.44], the leading mathematician in Berlin, who objected that projective geometry was not a metrical geometry: therefore, he claimed, it could not generate metrical concepts. However, Klein persisted and in a series of three papers, in 1871, 1872, and 1873, showed that all the known geometries could be regarded as subgeometries of projective geometry. His idea was to recast geometry as the study of a group acting on a space. Properties of figures (subsets of the space) that remain invariant under the action of the group are the geometric properties. So, for example, in a projective space of some dimension, the appropriate group for projective geometry is the group of all transformations that map lines to lines, and the subgroup that maps the interior of a given conic to itself may be regarded as the group of transformations of non-Euclidean geometry (see the [box](#) on p. 94). (For a fuller discussion of Klein's approach to geometry, see [I.3 §6].)

In the 1870s Klein's message was spread by the first and third of these papers, which were published in the recently founded journal *Mathematische Annalen*. As Klein's prestige grew, matters changed, and by the 1890s, when he had the second of the papers republished and translated into several languages, it was this, the *Erlangen Program*, that became well-known. It is named after the university where Klein became a professor, at the remarkably young age of twenty-three, but it was not his inaugural address. (That was about mathematics education.) For many years it was a singularly obscure publication, and it is unlikely that it had the effect on mathematics that some historians have come to suggest.

9 Convincing Others

Klein's work directed attention away from the *figures* in geometry and toward the *transformations* that do not alter the figures in crucial respects. For example, in Euclidean geometry the important transformations are the familiar rotations and translations (and reflections, if one chooses to allow them). These correspond to the motions of rigid bodies that contemporary psychologists saw as part of the way in which individuals learn the geometry of the space around them. But

this theory was philosophically contentious, especially when it could be extended to another metrical geometry, non-Euclidean geometry. Klein prudently entitled his main papers "On the so-called non-Euclidean geometry," to keep hostile philosophers at bay (in particular Lotze, who was the well-established Kantian philosopher at Göttingen). But with these papers and the previous work of Beltrami the case for non-Euclidean geometry was made, and almost all mathematicians were persuaded. They believed, that is, that alongside Euclidean geometry there now stood an equally valid mathematical system called non-Euclidean geometry. As for which one of these was true of space, it seemed so clear that Euclidean geometry was the sensible choice that there appears to have been little or no discussion. Lipschitz showed that it was possible to do all of mechanics in the new setting, and there the matter rested, a hypothetical case of some charm but no more. Helmholtz, the leading physicist of his day, became interested—he had known Riemann personally—and gave an account of what space would have to be if it was learned about through the free mobility of bodies. His first account was deeply flawed, because he was unaware of non-Euclidean geometry, but when Beltrami pointed this out to him he reworked it (in 1870). The reworked version also suffered from mathematical deficiencies, which were pointed out somewhat later by LIE [VI.53], but he had more immediate trouble from philosophers.

Their question was, "What sort of knowledge is this theory of non-Euclidean geometry?" Kantian philosophy was coming back into fashion, and in Kant's view knowledge of space was a fundamental pure a priori intuition, rather than a matter to be determined by experiment: without this intuition it would be impossible to have any knowledge of space at all. Faced with a rival theory, non-Euclidean geometry, neo-Kantian philosophers had a problem. They could agree that the mathematicians had produced a new and prolonged logical exercise, but could it be knowledge of the world? Surely the world could not have two kinds of geometry? Helmholtz hit back, arguing that knowledge of Euclidean geometry and non-Euclidean geometry would be acquired in the same way—through experience—but these empiricist overtones were unacceptable to the philosophers, and non-Euclidean geometry remained a problem for them until the early years of the twentieth century.

Mathematicians could not in fact have given a completely rigorous defense of what was becoming the

T&T note: in a perfect world figure 7 would appear on same spread as the box. Check at CRC stage.

Cross-ratios and distances in conics. A projective transformation of the plane sends four distinct points on a line, A, B, C, D , to four distinct collinear points, A', B', C', D' , in such a way that the quantity

$$\frac{AB}{AD} \frac{CD}{CB}$$

is preserved: that is,

$$\frac{AB}{AD} \frac{CD}{CB} = \frac{A'B'}{A'D'} \frac{C'D'}{C'B'}.$$

This quantity is called the *cross-ratio* of the four points A, B, C, D , and is written $CR(A, B, C, D)$.

In 1871, Klein described non-Euclidean geometry as the geometry of points inside a fixed conic, K , where the transformations allowed are the projec-

tive transformations that map K to itself and its interior to its interior (see figure 7). To define the distance between two points P and Q inside K , Klein noted that if the line PQ is extended to meet K at A and D , then the cross-ratio $CR(A, P, D, Q)$ does not change if one applies a projective transformation: that is, it is a *projective invariant*. Moreover, if R is a third point on the line PQ and the points lie in the order P, Q, R , then $CR(A, P, D, Q) CR(A, Q, D, R) = CR(A, P, D, R)$. Accordingly, he defined the distance between P and Q as $d(PQ) = -\frac{1}{2} \log CR(A, P, D, Q)$ (the factor of $-\frac{1}{2}$ is introduced to facilitate the later introduction of trigonometry). With this definition, distance is additive along a line: $d(PQ) + d(QR) = d(PR)$.

accepted position, but as the news spread that there were two possible descriptions of space, and that one could therefore no longer be certain that Euclidean geometry was correct, the educated public took up the question: what was the geometry of space? Among the first to grasp the problem in this new formulation was POINCARÉ [VI.61]. He came to mathematical fame in the early 1880s with a remarkable series of essays in which he reformulated Beltrami's disk model so as to make it *conformal*: that is, so that angles in non-Euclidean geometry were represented by the same angles in the model. He then used his new disk model to connect complex function theory, the theory of linear differential equations, RIEMANN SURFACE [III.81] theory, and non-Euclidean geometry to produce a rich new body of ideas. Then, in 1891, he pointed out that the disk model permitted one to show that any contradiction in non-Euclidean geometry would yield a contradiction in Euclidean geometry as well, and vice versa. Therefore, Euclidean geometry was consistent if and only if non-Euclidean geometry was consistent. A curious consequence of this was that if anybody *had* managed to derive the parallel postulate from the core of Euclidean geometry, then they would have inadvertently proved that Euclidean geometry was inconsistent!

One obvious way to try to decide which geometry described the actual universe was to appeal to physics. But Poincaré was not convinced by this. He argued in another paper (1902) that experience was open to many interpretations and there was no logical way of deciding what belonged to mathematics and what to physics. Imagine, for example, an elaborate set of measurements of angle sums of figures, perhaps on an astro-

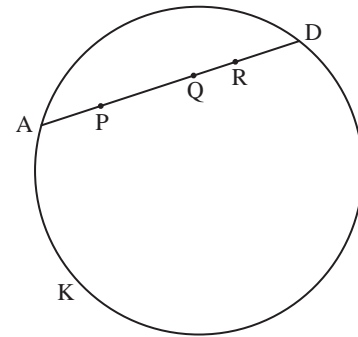


Figure 7 Three points, P, Q , and R , on a non-Euclidean straight line in Klein's projective model of non-Euclidean geometry.

nomical scale. Something would have to be taken to be straight, perhaps the paths of rays of light. Suppose, finally, that the conclusion is that the angle sum of a triangle is indeed less than two right angles by an amount proportional to the area of the triangle. Poincaré said that there were two possible conclusions: light rays are straight and the geometry of space is non-Euclidean; or light rays are somehow curved, and space is Euclidean. Moreover, he continued, there was no logical way to choose between these possibilities. All one could do was to make a convention and abide by it, and the sensible convention was to choose the simpler geometry: Euclidean geometry.

This philosophical position was to have a long life in the twentieth century under the name of *conventionalism*, but it was far from accepted in Poincaré's lifetime. A prominent critic of conventionalism was the

Italian Federigo Enriques, who, like Poincaré, was both a powerful mathematician and a writer of popular essays on issues in science and philosophy. He argued that one could decide whether a property was geometrical or physical by seeing whether we had any control over it. We cannot vary the law of gravity, but we can change the force of gravity at a point by moving matter around. Poincaré had compared his disk model to a metal disk that was hot in the center and got cooler as one moved outwards. He had shown that a simple law of cooling produced figures identical to those of non-Euclidean geometry. Enriques replied that heat was likewise something we can vary. A property such as Poincaré invoked, which was truly beyond our control, was not physical but geometric.

10 Looking Ahead

In the end, the question was not resolved in its own terms. Two developments moved mathematicians beyond the simple dichotomy posed by Poincaré. Starting in 1899, HILBERT [VI.63] began an extensive rewriting of geometry along axiomatic lines, which eclipsed earlier ideas of some Italian mathematicians and opened the way to axiomatic studies of many kinds. Hilbert's work captured very well the idea that if mathematics is sound, it is sound because of the nature of its reasoning, and led to profound investigations in mathematical logic. And in 1915 Einstein proposed his general theory of relativity, which is in large part a geometric theory of gravity. Confidence in mathematics was restored; our sense of geometry was much enlarged, and our insights into the relationships between geometry and space became considerably more sophisticated. Einstein made full use of contemporary ideas about geometry, and his achievement would have been unthinkable without Riemann's work. He described gravity as a kind of curvature in the four-dimensional manifold of spacetime (see GENERAL RELATIVITY AND THE EINSTEIN EQUATIONS [IV.17]). His work led to new ways of thinking about the large-scale structure of the universe and its ultimate fate, and to questions that remain unanswered to this day.

Further Reading

- Bonola, R. 1955. *History of Non-Euclidean Geometry*, translated by H. S. Carslaw and with a preface by F. Enriques. New York: Dover.
- Euclid. 1956. *The Thirteen Books of Euclid's Elements*, 2nd edn. New York: Dover.

- Gray, J. J. 1989. *Ideas of Space: Euclidean, Non-Euclidean, and Relativistic*, 2nd edn. Oxford: Oxford University Press.
- Gray, J. J. 2004. *Janos Bolyai, non-Euclidean Geometry and the Nature of Space*. Cambridge, MA: Burndy Library.
- Hilbert, D. 1899. *Grundlagen der Geometrie* (many subsequent editions). Tenth edn., 1971, translated by L. Unger, *Foundations of Geometry*. Chicago, IL: Open Court.
- Poincaré, H. 1891. Les géométries non-Euclidiennes. *Revue Générales des Sciences Pures et Appliquées* 2:769–74. (Reprinted, 1952, in *Science and Hypothesis*, pp. 35–50. New York: Dover.)
- . 1902. L'expérience et la géométrie. In *La Science et l'Hypothèse*, pp. 95–110. (Reprinted, 1952, in *Science and Hypothesis*, pp. 72–88. New York: Dover.)

II.3 The Development of Abstract Algebra

Karen Hunger Parshall

1 Introduction

What is algebra? To the high-school student encountering it for the first time, algebra is an unfamiliar abstract language of x 's and y 's, a 's and b 's, together with rules for manipulating them. These letters, some of them variables and some constants, can be used for many purposes. For example, one can use them to express straight lines as equations of the form $y = ax + b$, which can be graphed and thereby visualized in the Cartesian plane. Furthermore, by manipulating and interpreting these equations, it is possible to determine such things as what a given line's root is (if it has one)—that is, where it crosses the x -axis—and what its slope is—that is, how steep or flat it appears in the plane relative to the axis system. There are also techniques for solving simultaneous equations, or equivalently for determining when and where two lines intersect (or demonstrating that they are parallel).

Just when there already seem to be a lot of techniques and abstract manipulations involved in dealing with lines, the ante is upped. More complicated curves like quadratics, $y = ax^2 + bx + c$, and even cubics, $y = ax^3 + bx^2 + cx + d$, and quartics, $y = ax^4 + bx^3 + cx^2 + dx + e$, enter the picture, but the same sort of notation and rules apply, and similar sorts of questions are asked. Where are the roots of a given curve? Given two curves, where do they intersect?

Suppose now that the same high-school student, having mastered this sort of algebra, goes on to university and attends an algebra course there. Essentially gone

are the by now familiar x 's, y 's, a 's, and b 's; essentially gone are the nice graphs that provide a way to picture what is going on. The university course reflects some brave new world in which the algebra has somehow become "modern." This *modern* algebra involves abstract structures—GROUPS [I.3 §2.1], RINGS [III.83 §1], FIELDS [I.3 §2.2], and other so-called objects—each one defined in terms of a relatively small number of axioms and built up of substructures like subgroups, ideals, and subfields. There is a lot of moving around between these objects, too, via maps like group homomorphisms and ring AUTOMORPHISMS [I.3 §4.1]. One objective of this new type of algebra is to understand the underlying structure of the objects and, in doing so, to build entire theories of groups or rings or fields. These abstract theories may then be applied in diverse settings where the basic axioms are satisfied but where it may not be at all apparent a priori that a group or a ring or a field may be lurking. This, in fact, is one of modern algebra's great strengths: once we have proved a general fact about an algebraic structure, there is no need to prove that fact separately each time we come across an instance of that structure. This abstract approach allows us to recognize that contexts that may look quite different are in fact importantly similar.

How is it that two endeavors—the high-school analysis of polynomial equations and the modern algebra of the research mathematician—so seemingly different in their objectives, in their tools, and in their philosophical outlooks are both called "algebra"? Are they even related? In fact, they are, but the story of *how* they are is long and complicated.

2 Algebra before There Was Algebra: From Old Babylon to the Hellenistic Era

Solutions of what would today be recognized as first- and second-degree polynomial equations may be found in Old Babylonian cuneiform texts that date to the second millennium B.C.E. However, these problems were neither written in a notation that would be recognizable to our modern-day high-school student nor solved using the kinds of general techniques so characteristic of the high-school algebra classroom. Rather, particular problems were posed, and particular solutions obtained, from a series of recipe-like steps. No general theoretical justification was given, and the problems were largely cast geometrically, in terms of measurable line segments and surfaces of particular areas. Consider, for example, this problem, translated and transcribed from a clay tablet held in the British Museum

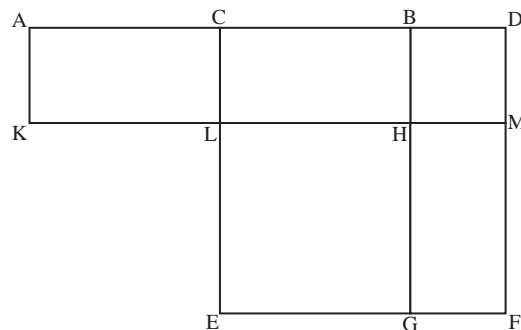


Figure 1 The sixth proposition from Euclid's Book II.

(catalogued as BM 13901, problem 1) that dates from between 1800 and 1600 B.C.E.:

The surface of my confrontation I have accumulated: 45' is it. 1, the projection, you posit. The moiety of 1 you break, 30' and 30' you make hold. 15' to 45' you append: by 1, 1 is equalside. 30' which you have made hold in the inside you tear out: 30' the confrontation.

This may be translated into modern notation as the equation $x^2 + 1x = \frac{3}{4}$, where it is important to notice that the Babylonian number system is base 60, so 45' denotes $\frac{45}{60} = \frac{3}{4}$. The text then lays out the following algorithm for solving the problem: take 1, the coefficient of the linear term, and halve it to get $\frac{1}{2}$. Square $\frac{1}{2}$ to get $\frac{1}{4}$. Add $\frac{1}{4}$ to $\frac{3}{4}$, the constant term, to get 1. This is the square of 1. Subtract from this the $\frac{1}{2}$ which you multiplied by to get $\frac{1}{2}$, the side of the square. The modern reader can easily see that this algorithm is equivalent to what is now called the quadratic formula, but the Babylonian tablet presents it in the context of a particular problem and repeats it in the contexts of other particular problems. There are no equations in the modern sense; the Babylonian writer is literally effecting a construction of plane figures. Similar problems and similar algorithmic solutions can also be found in ancient Egyptian texts such as the Rhind papyrus, believed to have been copied in 1650 B.C.E. from a text that was about a century and a half older.

The problem-oriented, untheoretical approach to mathematics characteristic of texts from this early period contrasts sharply with the axiomatic and deductive approach that EUCLID [VI.2] introduced into mathematics in around 300 B.C.E. in his magisterial, geometrical treatise, the *Elements*. (See GEOMETRY [II.2] for a further discussion of this work.) There, building on explicit definitions and a small number of axioms or self-evident truths, Euclid proceeded to deduce known—

and almost certainly some hitherto unknown—results within a strictly geometrical context. Geometry done in this axiomatic context defined Euclid’s standard of rigor. But what does this quintessentially geometrical text have to do with algebra? Consider the sixth proposition in Euclid’s Book II, ostensibly a book on plane figures, and in particular quadrilaterals:

If a straight line be bisected and a straight line be added to it in a straight line, the rectangle contained by the whole with the added straight line and the added straight line together with the square on the half is equal to the square on the straight line made up of the half and the added straight line.

While clearly a geometrical construction, it equally clearly describes two constructions—one a rectangle and one a square—that have equal areas. It therefore describes something that we should be able to write as an equation. Figure 1 gives the picture corresponding to Euclid’s construction: he proves that the area of rectangle ADMK equals the sum of rectangles CDML and HMFG. To do this, he adds the square on CB—namely, square LHGE—to CDML and HMFG. This gives square CDFE. It is not hard to see that this is equivalent to the high-school procedure of “completing the square” and to the algebraic equation $(2a + b)b + a^2 = (a + b)^2$, which we obtain by setting $CB = a$ and $BD = b$. Equivalent, yes, but for Euclid this is a specific *geometrical* construction and a particular *geometrical* equivalence. For this reason, he could not deal with anything but positive real quantities, since the *sides* of a geometrical figure could only be *measured* in those terms. Negative quantities did not and could not enter into Euclid’s fundamentally geometrical mathematical world. Nevertheless, in the historical literature, Euclid’s Book II has often been described as dealing with “geometrical algebra,” and, because of our easy translation of the book’s propositions into the language of algebra, it has been argued, albeit ahistorically, that Euclid *had* algebra but simply presented it geometrically.

Although Euclid’s geometrical standard of rigor came to be regarded as a pinnacle of mathematical achievement, it was in many ways not typical of the mathematics of classical Greek antiquity, a mathematics that focused less on systematization and more on the clever and individualistic solution of particular problems. There is perhaps no better exemplar of this than ARCHIMEDES [VI.3], held by many to have been one of the three or four greatest mathematicians of all time. Still, Archimedes, like Euclid, posed and solved

particular problems geometrically. As long as geometry defined the standard of rigor, not only negative numbers but also what we would recognize as polynomial equations of degree higher than three effectively fell outside the sphere of possible mathematical discussion. (As in the example from Euclid above, quadratic polynomials result from the geometrical process of completing the square; cubics could conceivably result from the geometrical process of completing the cube; but quartics and higher-degree polynomials could not be constructed in this way in familiar, three-dimensional space.) However, there was another mathematician of great importance to the present story, Diophantus of Alexandria (who was active in the middle of the third century C.E.). Like Archimedes, he posed particular problems, but he solved them in an algorithmic style much more reminiscent of the Old Babylonian texts than of Archimedes’ geometrical constructions, and as a result he was able to begin to exceed the bounds of geometry.

In his text *Arithmetica*, Diophantus put forward general, indeterminate problems, which he then restricted by specifying that the solutions should have particular forms, before providing specific solutions. He expressed these problems in a very different way from the purely rhetorical style that held sway for centuries after him. His notation was more algebraic and was ultimately to prove suggestive to sixteenth-century mathematicians (see below). In particular, he used special abbreviations that allowed him to deal with the first *six* positive and negative powers of the unknown as well as with the unknown to the zeroth power. Thus, whatever his mathematics was, it was not the “geometrical algebra” of Euclid and Archimedes.

Consider, for example, this problem from Book II of the *Arithmetica*: “To find three numbers such that the square of any one of them minus the next following gives a square.” In terms of modern notation, he began by restricting his attention to solutions of the form $(x + 1, 2x + 1, 4x + 1)$. It is easy to see that $(x + 1)^2 - (2x + 1) = x^2$ and $(2x + 1)^2 - (4x + 1) = 4x^2$, so two of the conditions of the problem are immediately satisfied, but he needed $(4x + 1)^2 - (x + 1) = 16x^2 + 7x$ to be a square as well. Arbitrarily setting $16x^2 + 7x = 25x^2$, Diophantus then determined that $x = \frac{7}{9}$ gave him what he needed, so a solution was $\frac{16}{9}, \frac{23}{9}, \frac{37}{9}$, and he was done. He provided no geometrical justification because in his view none was needed; a *single* numerical solution was all he required. He did not set up what we

would recognize as a more general set of equations and try to find all possible solutions.

Diophantus, who lived more than four centuries after Archimedes' death, was doing neither geometry nor algebra in our modern sense, yet the kinds of problems and the sorts of solutions he obtained for them were very different from those found in the works of either Euclid or Archimedes. The extent to which Diophantus created a wholly new approach, rather than drawing on an Alexandrian tradition of what might be called "algorithmic algebraic," as opposed to "geometric algebraic," scholarship is unknown. It is clear that by the time Diophantus's ideas were introduced into the Latin West in the sixteenth century, they suggested new possibilities to mathematicians long conditioned to the authority of geometry.

3 Algebra before There Was Algebra: The Medieval Islamic World

The transmission of mathematical ideas was, however, a complex process. After the fall of the Roman Empire and the subsequent decline of learning in the West, both the Euclidean and the Diophantine traditions ultimately made their way into the medieval Islamic world. There they were not only preserved—thanks to the active translation initiatives of Islamic scholars—but also studied and extended.

AL-KHWĀRIZMĪ [VI.5] was a scholar at the royally funded House of Wisdom in Baghdad. He linked the kinds of geometrical arguments Euclid had presented in Book II of his *Elements* with the indigenous problem-solving algorithms that dated back to Old Babylonian times. In particular, he wrote a book on practical mathematics, entitled *al-Kitāb al-mukhtaṣar fī ḥisāb al-jabr wa'l-muqābala* ("The compendious book on calculation by completion and balancing"), beginning it with a theoretical discussion of what we would now recognize as polynomial equations of the first and second degrees. (The latinization of the word "al-jabr" or "completion" in his title gave us our modern term "algebra.") Because he employed neither negative numbers nor zero coefficients, al-Khwārizmī provided a systematization in terms of six separate kinds of examples where we would need just one, namely $ax^2 + bx + c = 0$. He considered, for example, the case when "a square and 10 roots are equal to 39 units," and his algorithmic solution in terms of multiplications, additions, and subtractions was in precisely the same form as the above solution from tablet BM 13901. This, however,

was not enough for al-Khwārizmī. "It is necessary," he said, "that we should demonstrate geometrically the truth of the same problems which we have explained in numbers," and he proceeded to do this by "completing the square" in geometrical terms reminiscent of, but not as formal as, those Euclid used in Book II. (Abū Kāmil (ca. 850–930), an Egyptian Islamic mathematician of the generation after al-Khwārizmī, introduced a higher level of Euclidean formality into the geometric-algorithmic setting.) This juxtaposition made explicit how the relationships between geometrical areas and lines could be interpreted in terms of numerical multiplications, additions, and subtractions, a key step that would ultimately suggest a move away from the *geometrical* solution of *particular problems* and toward an *algebraic* solution of *general types of equations*.

Another step along this path was taken by the mathematician and poet Omar Khayyam (ca. 1050–1130) in a book he entitled *Al-jabr* after al-Khwārizmī's work. Here he proceeded to systematize and solve what we would recognize, in the absence of both negative numbers and zero coefficients, as the cases of the cubic equation. Following al-Khwārizmī, Khayyam provided geometrical justifications, yet his work, even more than that of his predecessor, may be seen as closer to a general problem-solving technique for specific cases of equations, that is, closer to the notion of algebra.

The Persian mathematician al-Karajī (who flourished in the early eleventh century) also knew well and appreciated the geometrical tradition stemming from Euclid's *Elements*. However, like Abū-Kāmil, he was aware of the Diophantine tradition too, and synthesized in more general terms some of the procedures Diophantus had laid out in the context of specific examples in the *Arithmetica*. Although Diophantus's ideas and style were known to these and other medieval Islamic mathematicians, they would remain unknown in the Latin West until their rediscovery and translation in the sixteenth century. Equally unknown in the Latin West were the accomplishments of Indian mathematicians, who had succeeded in solving some quadratic equations algorithmically by the beginning of the eighth century and who, like Brāhmagupta four hundred years later, had techniques for finding integer solutions to particular examples of what are today called Pell's equations, namely, equations of the form $ax^2 + b = y^2$, where a and b are integers and a is not a square.

4 Algebra before There Was Algebra: The Latin West

Concurrent with the rise of Islam in the East, the Latin West underwent a gradual cultural and political stabilization in the centuries following the fall of the Roman Empire. By the thirteenth century, this relative stability had resulted in the firm entrenchment of the Catholic Church as well as the establishment both of universities and of an active economy. Moreover, the Islamic conquest of most of the Iberian peninsula in the eighth century and the subsequent establishment there of an Islamic court, library, and research facility similar to the House of Wisdom in Baghdad brought the fruits of medieval Islamic scholarship to western Europe's doorstep. However, as Islam found its position on the Iberian peninsula increasingly compromised in the twelfth and thirteenth centuries, this Islamic learning, as well as some of the ancient Greek scholarship that the medieval Islamic scholars had preserved in Latin translation, began to filter into medieval Europe. In particular, FIBONACCI [VI.6], son of an influential administrator within the Pisan city state, encountered al-Khwārizmī's text and recognized not only the impact that the Arabic number system detailed there could have on accounting and commerce (Roman numerals and their cumbersome rules for manipulation were still widely in use) but also the importance of al-Khwārizmī's theoretical discussion, with its wedding of geometrical proof and the algorithmic solution of what we can interpret as first- and second-degree equations. In his 1202 book *Liber abbaci*, Fibonacci presented al-Khwārizmī's work almost verbatim, and extolled all of these virtues, thus effectively introducing this knowledge and approach into the Latin West.

Fibonacci's presentation, especially of the practical aspects of al-Khwārizmī's text, soon became well-known in Europe. So-called abacus schools (named after Fibonacci's text and not after the Chinese calculating instrument) sprang up all over the Italian peninsula, particularly in the fourteenth and fifteenth centuries, for the training of accountants and bookkeepers in an increasingly mercantilistic Western world. The teachers in these schools, the "maestri d'abaco," built on and extended the algorithms they found in Fibonacci's text. Another tradition, the Cossist tradition—after the German word "Coss" connoting algebra, that is, "Kunstrechnung" or "artful calculation"—developed simultaneously in the Germanic regions of Europe and aimed to introduce algebra into the mainstream there.

In 1494 the Italian Luca Pacioli published (by now this is the operative word: Pacioli's text is one of the earliest *printed* mathematical texts) a compendium of all known mathematics. By this time, the geometrical justifications that al-Khwārizmī and Fibonacci had presented had long since fallen from the mathematical vernacular. By reintroducing them in his book, the *Summa*, Pacioli brought them back to the mathematical fore. Not knowing of Khayyam's work, he asserted that solutions had been discovered only in the six cases treated by both al-Khwārizmī and Fibonacci, even though there had been abortive attempts to solve the cubic and even though he held out the hope that it could ultimately be solved.

Pacioli had highlighted a key unsolved problem: could algorithmic solutions be determined for the various cases of the cubic? And, if so, could these be justified geometrically with proofs similar in spirit to those found in the texts of al-Khwārizmī and Fibonacci?

Among several sixteenth-century Italian mathematicians who eventually managed to answer the first question in the affirmative was CARDANO [VI.7]. In his *Ars magna*, or *The Great Art*, of 1545, he presented algorithms with geometric justifications for the various cases of the cubic, effectively completing the cube where al-Khwārizmī and Fibonacci had completed the square. He also presented algorithms that had been discovered by his student Ludovico Ferrari (1522–65) for solving the cases of the quartic. These intrigued him, because, unlike the algorithms for the cubic, they were not justified geometrically. As he put it in his book, "all those matters up to and including the cubic are fully demonstrated, but the others which we will add, either by necessity or out of curiosity, we do not go beyond barely setting out." An algebra was breaking out of the geometrical shell in which it had been encased.

5 Algebra Is Born

This process was accelerated by the rediscovery and subsequent translation into Latin of Diophantus's *Arithmetica* in the 1560s, with its abbreviated presentational style and ungeometrical approach. Algebra, as a general problem-solving technique, applicable to questions in geometry, number theory, and other mathematical settings, was established in Raphael BOMBELLI's [VI.8] *Algebra* of 1572 and, more importantly, in VIÈTE's [VI.9] *In artem analyticem isagoge*, or *Introduction to the Analytic Art*, of 1591. The aim of the latter was, in Viète's words, "to leave no problem unsolved,"

and to this end he developed a true notation—using vowels to denote variables and consonants to denote coefficients—as well as methods for solving equations in one unknown. He called his techniques “specious logistics.”

Dimensionality—in the form of his so-called *law of homogeneity*—was, however, still an issue for Viète. As he put it, “[o]nly homogeneous magnitudes are to be compared to one another.” The problem was that he distinguished two types of magnitudes: “ladder magnitudes”—that is, variables (*A* side) (or x in our modern notation), (*A* square) (or x^2), (*A* cube) (or x^3), etc.; and “compared magnitudes”—that is, coefficients (*B* length) of dimension one, (*B* plane) of dimension two, (*B* solid) of dimension three, etc. In the light of his law of homogeneity, then, Viète could legitimately perform the operation (*A* cube) + (*B* plane)(*A* side) (or $x^3 + bx$ in our notation), since the dimension of (*A* cube) is three, as is that of the product of the two-dimensional coefficient (*B* plane) and the one-dimensional variable (*A* side), but he could not legally add the three-dimensional variable (*A* cube) to the two-dimensional product of the one-dimensional coefficient (*B* length) and the one-dimensional variable (*A* side) (or, again, $x^3 + bx$ in our notation). Be this as it may, his “analytic art” still allowed him to add, subtract, multiply, and divide *letters* as opposed to specific numbers, and those letters, as long as they satisfied the law of homogeneity, could be raised to the second, third, fourth, or, indeed, any power. He had a rudimentary algebra, although he failed to apply it to curves.

The first mathematicians to do that were FERMAT [VI.12] and DESCARTES [VI.11] in their independent development of the analytic geometry so familiar to the high-school algebra student of today. Fermat, and others like Thomas Harriot (ca. 1560–1621) in England, were influenced in their approaches by Viète, while Descartes not only introduced our present-day notational convention of representing variables by x ’s and y ’s and constants by a ’s, b ’s, and c ’s but also began the arithmetization of algebra. He introduced a unit that allowed him to interpret all geometrical magnitudes as line segments, whether they were x ’s, x^2 ’s, x^3 ’s, x^4 ’s, or any higher power of x , thereby removing concerns about homogeneity. Fermat’s main work in this direction was a 1636 manuscript written in Latin, entitled “Introduction to plane and solid loci” and circulated among the early seventeenth-century mathematical cognoscenti; Descartes’s was the *Geometry*, written in French as one of three appendices to his philo-

sophical tract, *Discourse on Method*, published in 1637. Both were regarded as establishing the identification of geometrical curves with equations in two unknowns, or in other words as establishing analytic geometry and thereby introducing *algebraic* techniques into the solution of what had previously been considered *geometrical* problems. In Fermat’s case, the curves were lines or conic sections—quadratic expressions in x and y ; Descartes did this too, but he also considered equations more generally, tackling questions about the roots of polynomial equations that were connected with transforming and reducing the polynomials.

In particular, although he gave no proof or even general statement of it, Descartes had a rudimentary version of what we would now call THE FUNDAMENTAL THEOREM OF ALGEBRA [V.15], the result that a polynomial equation $x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$ of degree n has precisely n roots over the field \mathbb{C} of complex numbers. For example, while he held that a given polynomial of degree n could be decomposed into n linear factors, he also recognized that the cubic $x^3 - 6x^2 + 13x - 10 = 0$ has three roots: the real root 2 and two complex roots. In his further exploration of these issues, moreover, he developed algebraic techniques, involving suitable transformations, for analyzing polynomial equations of the fifth and sixth degrees. Liberated from homogeneity concerns, Descartes was thus able to use his algebraic techniques freely to explore territory where the geometrically bound Cardano had clearly been reluctant to venture. NEWTON [VI.14] took the liberation of algebra from geometrical concerns a step further in his *Arithmetica universalis* (or *Universal Arithmetic*) of 1707, arguing for the complete arithmetization of algebra, that is, for modeling algebra and algebraic operations on the real numbers and the usual operations of arithmetic.

Descartes’s *Geometry* highlighted at least two problems for further algebraic exploration: the fundamental theorem of algebra and the solution of polynomial equations of degree greater than four. Although eighteenth-century mathematicians like D’ALEMBERT [VI.20] and EULER [VI.19] attempted proofs of the fundamental theorem of algebra, the first person to prove it rigorously was GAUSS [VI.26], who gave four distinct proofs over the course of his career. His first, an algebraic geometrical proof, appeared in his doctoral dissertation of 1799, while a second, fundamentally different proof was published in 1816, which in modern terminology essentially involved constructing the polynomial’s splitting field. While the fundamental theorem

of algebra established how many roots a given polynomial equation has, it did not provide insight into exactly what those roots were or how precisely to find them. That problem and its many mathematical repercussions exercised a number of mathematicians in the late eighteenth and nineteenth centuries and formed one of the strands of the mathematical thread that became modern algebra in the early twentieth century. Another emerged from attempts to understand the general behavior of systems of (one or more) polynomials in n unknowns, and yet another grew from efforts to approach number-theoretic questions algebraically.

6 The Search for the Roots of Algebraic Equations

The problem of finding roots of polynomials provides a direct link from the algebra of the high-school classroom to that of the modern research mathematician. Today's high-school student dutifully employs the quadratic formula to calculate the roots of second-degree polynomials. To derive this formula, one transforms the given polynomial into one that can be solved more easily. By more complicated manipulations of cubics and quartics, Cardano and Ferrari obtained formulas for the roots of those as well. It is natural to ask whether the same can be done for higher-degree polynomials. More precisely, are there formulas that involve just the usual operations of arithmetic—addition, subtraction, multiplication, and division—together with the extraction of roots? When there is such a formula, one says that the equation is *solvable by radicals*.

Although many eighteenth-century mathematicians (among them Euler, Alexandre-Théophile Vandermonde (1735–96), WARING [VI.21], and Étienne Bézout (1730–83)) contributed to the effort to decide whether higher-order polynomial equations are solvable by radicals, it was not until the years from roughly 1770 to 1830 that there were significant breakthroughs, particularly in the work of LAGRANGE [VI.22], ABEL [VI.33], and Gauss.

In a lengthy set of “Réflexions sur la résolution algébrique des équations” (Reflections on the algebraic resolution of equations) published in 1771, Lagrange tried to determine principles underlying the resolution of algebraic equations in general by analyzing in detail the specific cases of the cubic and the quartic. Building on the work of Cardano, Lagrange showed that a cubic of the form $x^3 + ax^2 + bx + c = 0$ could always be transformed into a cubic with no quadratic term

$x^3 + px + q = 0$ and that the roots of this could be written as $x = u + v$, where u^3 and v^3 are the roots of a certain quadratic polynomial equation. Lagrange was then able to show that if x_1, x_2, x_3 are the three roots of the cubic, the intermediate functions u and v could actually be written as $u = \frac{1}{3}(x_1 + \alpha x_2 + \alpha^2 x_3)$ and $v = \frac{1}{3}(x_1 + \alpha^2 x_2 + \alpha x_3)$, for α a primitive cube root of unity. That is, u and v could be written as rational expressions or resolvents in x_1, x_2, x_3 . Conversely, starting with a linear expression $y = Ax_1 + Bx_2 + Cx_3$ in the roots x_1, x_2, x_3 and then permuting the roots in all possible ways yielded six expressions each of which was a root of a particular sixth-degree polynomial equation. An analysis of the latter equation (which involved the exploitation of properties of symmetric polynomials) yielded the same expressions for u and v in terms of x_1, x_2, x_3 and the cube root of unity α . As Lagrange showed, this kind of two-pronged analysis—involving intermediate expressions rational in the roots that are solutions of a solvable equation as well as the behavior of certain rational expressions under permutation of the roots—yielded the complete solution in the cases both of the cubic and the quartic. It was *one* approach that encompassed the solution of *both* types of equation. But could this technique be extended to the case of the quintic and higher-degree polynomials? Lagrange was unable to push it through in the case of the quintic, but by building on his ideas, first his student Paolo Ruffini (1765–1822) at the turn of the nineteenth century and then, definitively, the young Norwegian mathematician Abel in the 1820s showed that, in fact, the quintic is *not* solvable by radicals. (See THE INSOLUBILITY OF THE QUINTIC [V.24].) This negative result, however, still left open the questions of which algebraic equations *were* solvable by radicals and why.

As Lagrange's analysis seemed to underscore, the answer to this question in the cases of the cubic and the quartic involved in a critical way the cube and fourth roots of unity, respectively. By definition, these satisfy the particularly simple polynomial equations $x^3 - 1 = 0$ and $x^4 - 1 = 0$, respectively. It was thus natural to examine the general case of the so-called cyclotomic equation $x^n - 1 = 0$ and ask for what values n the n th roots of unity are actually constructible. To put this question in equivalent algebraic terms: for which n is it possible to find a formula for the n th roots of unity that expresses them in terms of integers using the usual arithmetical operations and extraction of square (but not higher) roots? This was one of the many questions explored by Gauss in his wide-ranging, magiste-

rial, and groundbreaking 1801 treatise *Disquisitiones arithmeticae*. One of his most famous results was that the regular 17-gon (or, equivalently, a 17th root of unity) was constructible. In the course of his analysis, he not only employed techniques similar to those developed by Lagrange but also developed key concepts such as MODULAR ARITHMETIC [III.60] and the properties of the modular “worlds” \mathbb{Z}_p , for p a prime, and, more generally, \mathbb{Z}_n , for $n \in \mathbb{Z}^+$, as well as the notion of a primitive element (a generator) of what would later be termed a cyclic group.

Although it is not clear how well he knew Gauss’s work, in the years around 1830 GALOIS [VI.41] drew from the ideas both of Lagrange on the analysis of resolvents and of CAUCHY [VI.29] on permutations and substitutions to obtain a solution to the general problem of solvability of polynomial equations by radicals. Although his approach borrowed from earlier ideas, it was in one important respect fundamentally new. Whereas prior efforts had aimed at deriving an *explicit algorithm for calculating* the roots of a polynomial of a given degree, Galois formulated a theoretical process based on constructs more general than but derived from the given equation that allowed him to *assess whether or not that equation was solvable*.

To be more precise, Galois recast the problem into one in terms of two new concepts: fields (which he called “domains of rationality”) and groups (or, more precisely, groups of substitutions). A polynomial equation $f(x) = 0$ of degree n was reducible over its domain of rationality—the ground field from which its coefficients were taken—if all n of its roots were in that ground field; otherwise, it was irreducible over that field. It could, however, be reducible over some larger field. Consider, for example, the polynomial $x^2 + 1$ as a polynomial over \mathbb{R} , the field of real numbers. While we know from high-school algebra that this polynomial does not factor into a product of two real, linear factors (that is, there are no real numbers r_1 and r_2 such that $x^2 + 1 = (x - r_1)(x - r_2)$), it does factor over \mathbb{C} , the field of complex numbers, and, specifically, $x^2 + 1 = (x + \sqrt{-1})(x - \sqrt{-1})$. Thus, if we take all numbers of the form $a + b\sqrt{-1}$, where a and b belong to \mathbb{R} , then we enlarge \mathbb{R} to a new field \mathbb{C} in which the polynomial $x^2 + 1$ is reducible. If \mathbb{F} is a field and x is an element of \mathbb{F} that does not have an n th root in \mathbb{F} , then by a similar process we can adjoin an element y to \mathbb{F} and stipulate that $y^n = x$. We call y a *radical*. The set of all polynomial expressions in y , with coefficients in \mathbb{F} , can be shown to form a larger field. Galois showed that

if it was possible to enlarge \mathbb{F} by successively adjoining radicals to obtain a field K in which $f(x)$ factored into n linear factors, then $f(x) = 0$ was solvable by radicals. He developed a process that hinged both on the notion of adjoining an element—in particular, a so-called primitive element—to a given ground field and on the idea of analyzing the internal structure of this new, enlarged field via an analysis of the (finite) group of substitutions (automorphisms of K) that leave invariant all rational relations of the n roots of $f(x) = 0$. The group-theoretic aspects of Galois’s analysis were particularly potent; he introduced the notions, although not the modern terminology, of a normal subgroup of a group, a factor group, and a solvable group. Galois thus resolved the concrete problem of determining when a polynomial equation was solvable by radicals by examining it from the abstract perspective of groups and their internal structure.

Galois’s ideas, although sketched in the early 1830s, did not begin to enter into the broader mathematical consciousness until their publication in 1846 in LIOUVILLE’s [VI.39] *Journal des Mathématiques Pures et Appliquées*, and they were not fully appreciated until two decades later when first Joseph Serret (1819–85) and then JORDAN [VI.52] fleshed them out more fully. In particular, Jordan’s *Traité des substitutions et des équations algébriques* (“Treatise on substitutions and on algebraic equations”) of 1870 not only highlighted Galois’s work on the solution of algebraic equations but also developed the general structure theory of permutation groups as it had evolved at the hands of Lagrange, Gauss, Cauchy, Galois, and others. By the end of the nineteenth century, this line of development of group theory, stemming from efforts to solve algebraic equations by radicals, had intertwined with three others: the abstract notion of a group defined in terms of a group multiplication table, which was formulated by CAYLEY [VI.46], the structural work of mathematicians like Ludwig Sylow (1832–1918) and Otto Hölder (1859–1937), and the geometrical work of LIE [VI.53] and KLEIN [VI.57]. By 1893, when Heinrich Weber (1842–1914) codified much of this earlier work by giving the first actual abstract definitions of the notions both of group and field, thereby recasting them in a form much more familiar to the modern mathematician, groups and fields had been shown to be of central importance in a wide variety of areas, both mathematical and physical.

PUP: I can confirm that ‘ $x - r_1$ ’ is indeed correct here.

7 Exploring the Behavior of Polynomials in n Unknowns

The problem of solving algebraic equations involved finding the roots of polynomials in *one* unknown. At least as early as the late seventeenth century, however, mathematicians like LEIBNIZ [VI.15] had been interested in techniques for solving simultaneously systems of linear equations in more than two variables. Although his work remained unknown at the time, Leibniz considered three linear equations in three unknowns and determined their simultaneous solvability based on the value of a particular expression in the coefficients of the system. This expression, equivalent to what Cauchy would later call the DETERMINANT [III.15] and which would ultimately be associated with an $n \times n$ square array or MATRIX [I.3 §4.2] of coefficients, was also developed and analyzed independently by Gabriel Cramer (1704–52) in the mid eighteenth century in the general context of the simultaneous solution of a system of n linear equations in n unknowns. From these beginnings, a theory of determinants, independent of the context of solving systems of linear equations, quickly became a topic of algebraic study in its own right, attracting the attention of Vandermonde, LAPLACE [VI.23], and Cauchy, among others. Determinants were thus an example of a new algebraic construct, the properties of which were then systematically explored.

Although determinants came to be viewed in terms of what SYLVESTER [VI.42] would dub matrices, a theory of matrices proper grew initially from the context not of solving simultaneous linear equations but rather of linearly transforming the variables of homogeneous polynomials in two, three, or more generally n variables. In the *Disquisitiones arithmeticae*, for example, Gauss considered how binary and ternary quadratic forms with integer coefficients—expressions of the form $a_1x^2 + 2a_2xy + a_3y^2$ and $a_1x^2 + a_2y^2 + a_3z^2 + 2a_4xz + 2a_5yz$, respectively—are affected by a linear transformation of their variables. In the ternary case, he applied the linear transformation $x = \alpha x' + \beta y' + \gamma z'$, $y = \alpha' x' + \beta' y' + \gamma' z'$, and $z = \alpha'' x' + \beta'' y' + \gamma'' z'$ to derive a new ternary form. He denoted the linear transformation of the variables by the square array

$$\begin{array}{ccc} \alpha, & \beta, & \gamma \\ \alpha', & \beta', & \gamma' \\ \alpha'', & \beta'', & \gamma'' \end{array}$$

and, in showing what the composition of two such transformations was, gave an explicit example of matrix multiplication. By the middle of the nineteenth century, Cayley had begun to explore matrices per se and had established many of the properties that the theory of matrices as a mathematical system in its own right enjoys. This line of algebraic thought was eventually reinterpreted in terms of the theory of algebras (see below) and developed into the independent area of linear algebra and the theory of VECTOR SPACES [I.3 §2.3].

Another theory that arose out of the analysis of linear transformations of homogeneous polynomials was the theory of invariants, and this too has its origins in some sense in Gauss's *Disquisitiones*. As in his study of ternary quadratic forms, Gauss began his study of binary forms by applying a linear transformation, specifically, $x = \alpha x' + \beta y'$, $y = \gamma x' + \delta y'$. The result was the new binary form $a'_1(x')^2 + 2a'_2x'y' + a'_3(y')^2$, where, explicitly, $a'_1 = a_1\alpha^2 + 2a_2\alpha\gamma + a_3\gamma^2$, $a'_2 = a_1\alpha\beta + a_2(\alpha\delta + \beta\gamma) + a_3\gamma\delta$, and $a'_3 = a_1\beta^2 + 2a_2\beta\delta + a_3\delta^2$. As Gauss noted, if you multiply the second of these equations by itself and subtract from this the product of the first and the third equations, you obtain the relation $a'^2_2 - a'_1a'_3 = (a^2_2 - a_1a_3)(\alpha\delta - \beta\gamma)^2$. To use language that Sylvester would develop in the early 1850s, Gauss realized that the expression $a^2_2 - a_1a_3$ in the coefficients of the original binary quadratic form is an *invariant* in the sense that it remains unchanged up to a power of the determinant of the linear transformation. By the time Sylvester coined the term, the invariant phenomenon had also appeared in the work of the English mathematician BOOLE [VI.43], and had attracted Cayley's attention. It was not until after Cayley and Sylvester met in the late 1840s, however, that the two of them began to pursue a theory of invariants proper, which aimed to determine all invariants for homogeneous polynomials of degree m in n unknowns as well as simultaneous invariants for systems of such polynomials.

Although Cayley and (especially) Sylvester pursued this line of research from a purely algebraic point of view, invariant theory also had number-theoretic and geometric implications, the former explored by Gotthold Eisenstein (1823–52) and HERMITE [VI.47], the latter by Otto Hesse (1811–74), Paul Gordan (1837–1912), and Alfred Clebsch (1833–72), among others. It was of particular interest to understand how many “genuinely distinct” invariants were associated with a specific form, or system of forms. In 1868, Gordan

achieved a fundamental breakthrough by showing that the invariants associated with any binary form in n variables can always be expressed in terms of a finite number of them. By the late 1880s and early 1890s, however, HILBERT [VI.63] brought new, abstract concepts associated with the theory of algebras (see below) to bear on invariant theory and, in so doing, not only reproved Gordan's result but also showed that the result was true for forms of degree m in n unknowns. With Hilbert's work, the emphasis shifted from the concrete calculations of his English and German predecessors to the kind of structurally oriented existence theorems that would soon be associated with abstract, modern algebra.

8 The Quest to Understand the Properties of "Numbers"

As early as the sixth century B.C.E., the Pythagoreans had studied the properties of numbers formally. For example, they defined the concept of a *perfect number*, which is a positive integer, such as $6 = 1 + 2 + 3$ and $28 = 1 + 2 + 4 + 7 + 14$, which is the sum of its divisors (excluding the integer itself). In the sixteenth century, Cardano and Bombelli had willingly worked with new expressions, complex numbers, of the form $a + \sqrt{-b}$, for real numbers a and b , and had explored their computational properties. In the seventeenth century, Fermat famously claimed that he could prove that the equation $x^n + y^n = z^n$, for n an integer greater than 2, had no solutions in the integers, except for the trivial cases when $z = x$ or $z = y$ and the remaining variable is zero. The latter result, known as FERMAT'S LAST THEOREM [V.12], generated many new ideas, especially in the eighteenth and nineteenth centuries, as mathematicians worked to find an actual proof of Fermat's claim. Central to their efforts were the creation and algebraic analysis of new types of number systems that extended the integers in much the same way that Galois had extended fields. This flexibility to create and analyze new number systems was to become one of the hallmarks of modern algebra as it would develop into the twentieth century.

One of the first to venture down this path was Euler. In the proof of Fermat's last theorem for the $n = 3$ case that he gave in his *Elements of Algebra* of 1770, Euler introduced the system of numbers of the form $a + b\sqrt{-3}$, where a and b are integers. He then blithely proceeded to factorize them into primes, without further justification, just as he would have factorized

ordinary integers. By the 1820s and 1830s, Gauss had launched a more systematic study of numbers that are now called the *Gaussian integers*. These are all numbers of the form $a + b\sqrt{-1}$, for integers a and b . He showed that, like the integers, the Gaussian integers are closed under addition, subtraction, and multiplication; he defined the notions of unit, prime, and norm in order to prove an analogue of THE FUNDAMENTAL THEOREM OF ARITHMETIC [V.16] for them. He thereby demonstrated that there were whole new algebraic worlds to create and explore. (See ALGEBRAIC NUMBERS [IV.3] for more on these topics.)

Whereas Euler had been motivated in his work by Fermat's last theorem, Gauss was trying to generalize the LAW OF QUADRATIC RECIPROCITY [V.30] to a law of biquadratic reciprocity. In the quadratic case, the problem was the following. If a and m are integers with $m \geq 2$, then we say that a is a *quadratic residue mod m* if the equation $x^2 = a$ has a solution mod m ; that is, if there is an integer x such that x^2 is congruent to a mod m . Now suppose that p and q are distinct odd primes. If you know whether p is a quadratic residue mod q , is there a simple way of telling whether q is a quadratic residue mod p ? In 1785, Legendre had posed and answered this question—the status of q mod p will be the same as that of p mod q if at least one of p and q is congruent to 1 mod 4, and different if they are both congruent to 3 mod 4—but he had given a faulty proof. By 1796, Gauss had come up with the first rigorous proof of the theorem (he would ultimately give eight different proofs of it), and by the 1820s he was asking the analogous question for the case of two biquadratic equivalences $x^4 \equiv p \pmod{q}$ and $y^4 \equiv q \pmod{p}$. It was in his attempts to answer this new question that he introduced the Gaussian integers and signaled at the same time that the theory of residues of higher degrees would make it necessary to create and analyze still other new sorts of "integers." Although Eisenstein, DIRICHLET [VI.36], Hermite, KUMMER [VI.40], and KRONECKER [VI.48], among others, pushed these ideas forward in this Gaussian spirit, it was DEDEKIND [VI.50] in his tenth supplement to Dirichlet's *Vorlesungen über Zahlentheorie* (*Lectures on Number Theory*) of 1871 who fundamentally reconceptualized the problem by treating it not number theoretically but rather set theoretically and axiomatically. Dedekind introduced, for example, the general notions—if not what would become the precise axiomatic definitions—of fields, rings, IDEALS [III.83 §2], and MODULES [III.83 §3] and analyzed his number-theoretic setting in terms of

these new, abstract constructs. His strategy was, from a philosophical point of view, not unlike that of Galois: translate the “concrete” problem at hand into new, more abstract terms in order to solve it more cleanly at a “higher” level. In the early twentieth century, NOETHER [VI.76] and her students, among them Bartel van der Waerden (1903–96), would develop Dedekind’s ideas further to help create the structural approach to algebra so characteristic of the twentieth century.

Parallel to this nineteenth-century, number-theoretic evolution of the notion of “number” on the continent of Europe, a very different set of developments was taking place, initially in the British Isles. From the late eighteenth century, British mathematicians had debated not only the nature of number—questions such as, “Do negative and imaginary numbers make sense?”—but also the meaning of algebra—questions like, “In an expression like $ax + by$, what values may a , b , x , and y legitimately take on and what precisely may ‘+’ connote?” By the 1830s, the Irish mathematician HAMILTON [VI.37] had come up with a “unified” interpretation of the complex numbers that circumvented, in his view, the logical problem of adding a real number and an imaginary one, an apple and an orange. Given real numbers a and b , Hamilton conceived of the complex number $a + b\sqrt{-1}$ as the ordered pair (he called it a “couple”) (a, b) . He then defined addition, subtraction, multiplication, and division of such couples. As he realized, this also provided a way of representing numbers in the complex plane, and so he naturally asked whether he could construct algebraic, ordered triples so as to represent points in 3-space. After a decade of contemplating this question off and on, Hamilton finally answered it not for triples but for quadruples, the so-called QUATERNIONS [III.78], “numbers” of the form $(a, b, c, d) := a + bi + cj + dk$, where a , b , c , and d are real and where i , j , k satisfy the relations $ij = -ji = k$, $jk = -kj = i$, $ki = -ik = j$, $i^2 = j^2 = k^2 = -1$. As in the two-dimensional case, addition is defined component-wise, but multiplication, while definable in such a way that every nonzero element has a multiplicative inverse, is not commutative. Thus, this new number system did not obey all of the “usual” laws of arithmetic.

Although some of Hamilton’s British contemporaries questioned the extent to which mathematicians were free to create such new mathematical worlds, others, like Cayley, immediately took the idea further and created a system of ordered 8-tuples, the octonions, the multiplication of which was neither commutative nor even, as was later discovered, associa-

tive. Several questions naturally arise about such systems, but one that Hamilton asked was what happens if the field of coefficients, the base field, is not the reals but rather the complexes? In that case, it is easy to see that the product of the two nonzero complex quaternions $(-\sqrt{-1}, 0, 1, 0) = -\sqrt{-1} + j$ and $(\sqrt{-1}, 0, 1, 0) = \sqrt{-1} + j$ is $1 + j^2 = 1 + (-1) = 0$. In other words, the complex quaternions contain zero divisors—nonzero elements the product of which is zero—another phenomenon that distinguishes their behavior fundamentally from that of the integers. As it flourished in the hands of mathematicians like Benjamin Peirce (1809–80), FROBENIUS [VI.58], Georg Scheffers (1866–1945), Theodor Molien (1861–1941), CARTAN [VI.69], and Joseph H. M. Wedderburn (1882–1948), among others, this line of thought resulted in a freestanding theory of algebras. This naturally intertwined with developments in the theory of matrices (the $n \times n$ matrices form an algebra of dimension n^2 over their base field) as it had evolved through the work of Gauss, Cayley, and Sylvester. It also merged with the not unrelated theory of n -dimensional vector spaces (n -dimensional algebras are n -dimensional vector spaces with a vector multiplication as well as a vector addition and scalar multiplication) that issued from ideas like those of Hermann Grassmann (1809–77).

9 Modern Algebra

By 1900, many new algebraic structures had been identified and their properties explored. Structures that were first isolated in one context were then found to appear, sometimes unexpectedly, in others: thus, these new structures were mathematically more general than the problems that had led to their discovery. In the opening decades of the twentieth century, algebraists (the term is not ahistorical by 1900) increasingly recognized these commonalities—these shared structures such as groups, fields and rings—and asked questions at a more abstract level. For example, what are all of the finite simple groups? Can they be classified? (See THE CLASSIFICATION OF FINITE SIMPLE GROUPS [V.8].) Moreover, inspired by the set-theoretic and axiomatic work of CANTOR [VI.54], Hilbert, and others, they came to appreciate the common standard of analysis and comparison that axiomatization could provide. Coming from this axiomatic point of view, Ernst Steinitz (1871–1928), for example, laid the groundwork for an abstract theory of fields in 1910, while Abraham Fraenkel (1891–1965) did the same for an abstract theory of rings four

years later. As van der Waerden came to realize in the late 1920s, these developments could be interpreted as dovetailing philosophically with results like Hilbert's in invariant theory and Dedekind's and Noether's in the algebraic theory of numbers. That interpretation, laid out in 1930 in van der Waerden's classic textbook *Mod-erne Algebra*, codified the structurally oriented "modern algebra" that subsumed the algebra of polynomials of the high-school classroom and that continues to characterize algebraic thought today.

Further Reading

- Bashmakova, I., and G. Smirnova. 2000. *The Beginnings and Evolution of Algebra*, translated by A. Shenitzer. Washington, DC: The Mathematical Association of America.
- Corry, L. 1996. *Modern Algebra and the Rise of Mathematical Structures*. Science Networks, volume 17. Basel: Birkhäuser.
- Edwards, H. M. 1984. *Galois Theory*. New York: Springer.
- Heath, T. L. 1956. *The Thirteen Books of Euclid's Elements*, 2nd edn. (3 vols.). New York: Dover.
- Høyrup, J. 2002. *Lengths, Widths, Surfaces: A Portrait of Old Babylonian Algebra and Its Kin*. New York: Springer.
- Klein, J. 1968. *Greek Mathematical Thought and the Origin of Algebra*, translated by E. Brann. Cambridge, MA: The MIT Press.
- Netz, R. 2004. *The Transformation of Mathematics in the Early Mediterranean World: From Problems to Equations*. Cambridge: Cambridge University Press.
- Parshall, K. H. 1988. The art of algebra from al-Khwārizmī to Viète: A study in the natural selection of ideas. *History of Science* 26:129–64.
- . 1989. Toward a history of nineteenth-century invariant theory. In *The History of Modern Mathematics*, edited by D. E. Rowe and J. McCleary, volume 1, pp. 157–206. Amsterdam: Academic Press.
- Sesiano, J. 1999. *Une Introduction à l'histoire de l'algèbre: Résolution des équations des Mésopotamiens à la Renaissance*. Lausanne: Presses Polytechniques et Universitaires Romandes.
- Van der Waerden, B. 1985. *A History of Algebra from al-Khwārizmī to Emmy Noether*. New York: Springer.
- Wussing, H. 1984. *The Genesis of the Abstract Group Concept: A Contribution to the History of the Origin of Abstract Group Theory*, translated by A. Shenitzer. Cambridge, MA: The MIT Press.

II.4 Algorithms

Jean-Luc Chabert

1 What Is an Algorithm?

It is not easy to give a precise definition of the word "algorithm." One can provide approximate synonyms:

some other words that (sometimes) mean roughly the same thing are "rule," "technique," "procedure," and "method." One can also give good examples, such as long multiplication, the method one learns in high school for multiplying two positive integers together. However, although informal explanations and well-chosen examples do give a good idea of what an algorithm is, the concept has undergone a long evolution: it was not until the twentieth century that a satisfactory formal definition was achieved, and ideas about algorithms have evolved further even since then. In this article, we shall try to explain some of these developments and clarify the contemporary meaning of the term.

1.1 Abacists and Algorithmists

Returning to the example of multiplication, an obvious point is that how you try to multiply two numbers together is strongly influenced by how you represent those numbers. To see this, try multiplying the Roman numerals CXLVII and XXIX together without first converting them into their decimal counterparts, 147 and 29. It is difficult and time-consuming, and explains why arithmetic in the Roman empire was extremely rudimentary. A numeration system can be additive, as it was for the Romans, or *positional*, like ours today. If it is positional, then it can use one or several bases—for instance, the Sumerians used both base 10 and base 60.

For a long time, many processes of calculation used *abacuses*. Originally, these were lines traced on sand, onto which one placed stones (the Latin for small stone is *calculus*) to represent numbers. Later there were counting tables equipped with rows or columns onto which one placed tokens. These could be used to represent numbers to a given base. For example, if the base was 10, then a token would represent one unit, ten units, one hundred units, etc., according to which row or column it was in. The four arithmetic operations could then be carried out by moving the tokens according to precise rules. The Chinese counting frame can be regarded as a version of the abacus.

In the twelfth century, when the Arabic mathematical works were translated into Latin, the denary positional numeration system spread through Europe. This system was particularly suitable for carrying out the arithmetic operations, and led to new methods of calculation. The term *algorithmus* was introduced to refer to these, and to distinguish them from the traditional methods that used tokens on an abacus.

Although the signs for the numerals had been adapted from Indian practice, the numerals became

known as Arabic. And the origin of the word “algorithm” is Arabic: it arose from a distortion of the name AL-KHWĀRIZMĪ [VI.5], who was the author of the oldest known work on algebra, in the first half of the ninth century. His treatise, entitled *al-Kitāb al-mukhtaṣar fī ḥisāb al-jabr wa’l-muqābala* (“The compendious book on calculation by completion and balancing”), gave rise to the word “algebra.”

1.2 Finiteness

As we have just seen, in the Middle Ages the term “algorithm” referred to the processes of calculation based on the decimal notation for the integers. However, in the seventeenth century, according to D’ALEMBERT’s [VI.20] *Encyclopédie*, the word was used in a more general sense, referring not just to arithmetic but also to methods in algebra and to other calculational procedures such as “the algorithm of the integral calculus” or “the algorithm of sines.”

Gradually, the term came to mean any process of systematic calculation that could be carried out by means of very precise rules. Finally, with the growing role of computers, the important role of *finiteness* was fully understood: it is essential that the process stops and provides a result after a finite time. Thus one arrives at the following naive definition:

An algorithm is a set of finitely many rules for manipulating a finite amount of data in order to produce a result in a finite number of steps.

Note the insistence on finiteness: finiteness in the writing of the algorithm and finiteness in the implementation of the algorithm.

The formulation above is not of course a mathematical definition in the classical sense of the term. As we shall see later, it was important to formalize it further. But for now, let us be content with this “definition” and look at some classical examples of algorithms in mathematics.

2 Three Historical Examples

A feature of algorithms that we have not yet mentioned is *iteration*, or the repetition of simple procedures. To see why iteration is important, consider once again the example of long multiplication. This is a method that works for positive integers of any size. As the numbers get larger, the procedure takes longer, but—and this is of vital importance—the method is “the same”: if you understand how to multiply two three-digit numbers together, then you do not need to learn any new

principles in order to multiply two 137-digit numbers together (even if you might be rather reluctant to do the calculation). The reason for this is that the method for long multiplication involves a great deal of carefully structured repetition of much smaller tasks, such as multiplying two one-digit numbers together. We shall see that iteration plays a very important part in the algorithms to be discussed in this section.

2.1 Euclid’s Algorithm: Iteration

One of the best, and most often used, examples to illustrate the nature of algorithms is EUCLID’S ALGORITHM [III.22], which goes back to the third century B.C.E. It is a procedure described by EUCLID [VI.2] to determine the *greatest common divisor* (gcd) of two positive integers a and b . (Sometimes the greatest common divisor is known as the *highest common factor* (hcf).)

When one first meets the concept of the greatest common divisor of a and b , it is usually defined to be the largest positive integer that is a divisor (or factor) of both a and b . However, for many purposes it is more convenient to think of it as the unique positive integer d with the following two properties. First, d is a divisor of a and b , and second, if c is any other divisor of a and b , then d is divisible by c . The method for determining d is provided by the first two propositions of Book VII of Euclid’s *Elements*. Here is the first one: “Two unequal numbers being set out, and the less being continually subtracted in turn from the greater, if the number which is left never measures the one before it until a unit is left, the original numbers will be prime to one another.” In other words, if by carrying out successive alternate subtractions one obtains the number 1, then the gcd of the two numbers is equal to 1. In this case one says that the numbers are *relatively prime* or *coprime*.

2.1.1 Alternate Subtractions

Let us describe Euclid’s procedure in general. It is based on two simple observations:

- (i) if $a = b$ then the gcd of a and b is b (or a);
- (ii) d is a common divisor of a and b if and only if it is a common divisor of $a - b$ and b , which implies that the gcd of a and b is the same as the gcd of $a - b$ and b .

Now suppose that we wish to determine the gcd of a and b and suppose that $a \geq b$. If $a = b$ then obser-

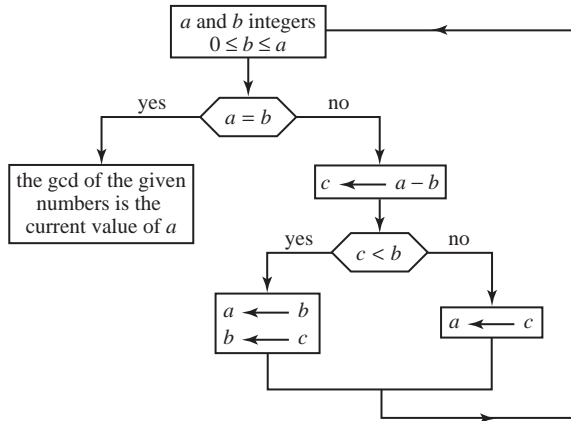


Figure 1 A flow chart for the procedure in Euclid's algorithm.

vation (i) tells us that the gcd is b . Otherwise, observation (ii) tells us that the answer will be the same as it is for the two numbers $a - b$ and b . If we now let a_1 be the larger of these two numbers and b_1 the smaller (of course, if they are equal then we just set $a_1 = b_1 = b$), then we are faced with the same task that we started with—to determine the gcd of two numbers—but the larger of these two numbers, a_1 , is smaller than a , the larger of the original two numbers. We can therefore repeat the process: if $a_1 = b_1$ then the gcd of a_1 and b_1 , and hence that of a and b , is b_1 , and otherwise we replace a_1 by $a_1 - b_1$ and reorganize the numbers $a_1 - b_1$ and b_1 so that if one of them is larger then it comes first.

One further observation is needed if we want to show that this procedure works. It is the following fundamental fact about the positive integers, sometimes known as the *well-ordering principle*.

- (iii) A strictly decreasing sequence of positive integers $a_0 > a_1 > a_2 > \dots$ must be finite.

Since the iterative procedure just described produces exactly such a strictly decreasing sequence, the iterations must eventually stop, which means that at some point a_k and b_k will be equal, and that value is thus the gcd of a and b (see figure 1).

2.1.2 Euclidean Divisions

Euclid's algorithm is usually described in a slightly different way. One makes use of a more complex procedure called *Euclidean division*—that is, division with remainder—which greatly reduces the number of steps

that the algorithm takes. The basic fact underlying this procedure is that if a and b are two positive integers then there are (unique) integers q and r such that

$$a = bq + r \quad \text{and} \quad 0 \leq r < b.$$

The number q is called the *quotient* and r is the *remainder*. Remarks (i) and (ii) above are then replaced by the following ones:

- (i') if $r = 0$ then the gcd of a and b is equal to b ;
(ii') the gcd of a and b is the same as the gcd of b and r .

This time, at the first step, one replaces (a, b) by (b, r) . If $r \neq 0$, then at the second step one replaces (b, r) by (r, r_1) , where r_1 is the remainder in the division of b by r , and so on. The sequence of remainders is strictly decreasing ($b > r > r_1 > r_2 \geq 0$), so the process stops and the gcd is the last nonzero remainder.

It is not hard to see that the two approaches are equivalent. Suppose, for example, that $a = 103\,438$ and $b = 37$. If you use the first approach, then you will repeatedly subtract 37 from 103 438 until you reach a number that is smaller than 37. This number will be the remainder when 103 438 is divided by 37, which is the first number you would calculate if you used the second approach. Thus, the reason for the second approach is that repeated subtraction can be a very inefficient way of calculating remainders. This efficiency gain is very important in practice: the second approach gives rise to a **POLYNOMIAL-TIME ALGORITHM** [IV.21 §2], while the time taken by the first is exponentially long.

2.1.3 Generalizations

Euclid's algorithm can be generalized to many other contexts where we have notions of addition, subtraction, and multiplication. For example, there is a variant of it that applies to the RING [III.83 §1] $\mathbb{Z}[i]$ of *Gaussian integers*, that is, numbers of the form $a + bi$, where a and b are ordinary integers. It can also be applied to the ring of all polynomials with real coefficients (or coefficients in any field, for that matter). The one requirement is that we should be able to find some analogue of the notion of division with remainder, after which the algorithm is virtually identical to the algorithm for positive integers. For example, we have the following statement for polynomials: given any two polynomials A and B with B not the zero polynomial, there are polynomials Q and R such that $A = BQ + R$ and either $R = 0$ or the degree of R is less than the degree of B .

As Euclid noticed (*Elements*, Book X, proposition 2), one may also carry out the procedure on pairs of numbers a and b that are not necessarily integers. It is easy to check that the process will stop if and only if the ratio a/b is a rational number. This observation leads to the concept of CONTINUED FRACTIONS [III.22], which are discussed in part III. They were not studied explicitly before the seventeenth century, but the roots of the idea can be traced back to ARCHIMEDES [VI.3].

2.2 The Method of Archimedes to Calculate π : Approximation and Finiteness

The ratio of the circumference of a circle to the diameter is a constant that has been denoted by π since the eighteenth century (see the article “ π ” in part III). Let us see how Archimedes, in the third century B.C.E., obtained the classical approximation $\frac{22}{7}$ for this ratio. If one draws inscribed polygons (whose vertices lie on the circle) and circumscribed polygons (whose sides are tangent to the circle) and if one computes the length of these polygons, then one obtains lower and upper bounds for the value of π , since the circumference of the circle is greater than the length of any inscribed polygon and less than the length of any circumscribed polygon (figure 2). Archimedes started with regular hexagons, and then repeatedly doubled the number of sides, obtaining more and more precise bounds. He finished with ninety-six-sided polygons, obtaining the estimates

$$3 + \frac{10}{71} \leq \pi \leq 3 + \frac{1}{7}.$$

This process clearly involves iteration, but is it right to call it an algorithm? Strictly speaking it is not: however many sides you take for your polygon, all you will get is an approximation to π , so the process is not finite. However, what we do have is an algorithm that will calculate π to any desired accuracy: for example, if you demand an approximation that is correct to ten decimal places, then after a finite number of steps the algorithm will give you one. What matters now is that the process *converges*. That is, it is important that the values that come out of the iteration get arbitrarily close to π . The geometric origin of the method can be used to prove that this is indeed the case, and in 1609 in Germany Ludolph van Ceulen obtained an approximation accurate to thirty-five decimal places using polygons with 2^{62} sides.

Nevertheless, there is a clear difference between this algorithm for approximating π and Euclid’s algorithm

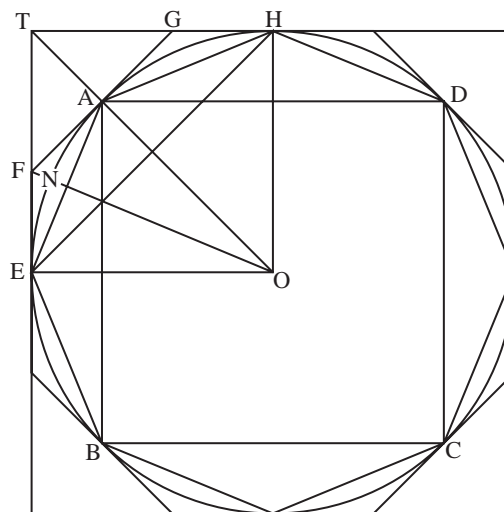


Figure 2 Approximation of π .

for calculating the gcd of two positive integers. Algorithms like Euclid’s are often called *discrete algorithms*, and are contrasted with *numerical algorithms*, which are algorithms that are used to compute numbers that are not integers (see NUMERICAL ANALYSIS [IV.20]).

2.3 The Newton-Raphson Method: Recurrence Formulas

In around 1670, NEWTON [VI.14] devised a method for finding roots of equations, which he explained with reference to the example $x^3 - 2x - 5 = 0$. His explanation starts with the observation that the root x is approximately equal to 2. He therefore writes $x = 2 + p$ and obtains an equation for p by substituting $2 + p$ for x in the original equation. This new equation works out to be $p^3 + 6p^2 + 10p - 1 = 0$. Because x is close to 2, p is small, so he then estimates p by forgetting the terms p^3 and $6p^2$ (since these should be considerably smaller than $10p - 1$). This gives him the equation $10p - 1 = 0$, or $p = \frac{1}{10}$. Of course, this is not an exact solution, but it provides him with a new and better approximation, 2.1, for x . He then repeats the process, writing $x = 2.1 + q$, substituting to obtain an equation for q , solving this equation approximately, and refining his estimate still further. The estimate he obtains for q is -0.0054 , so the next approximation for x is 2.0946.

How, though, can we be sure that this process really does converge to x ? Let us examine the method more closely.

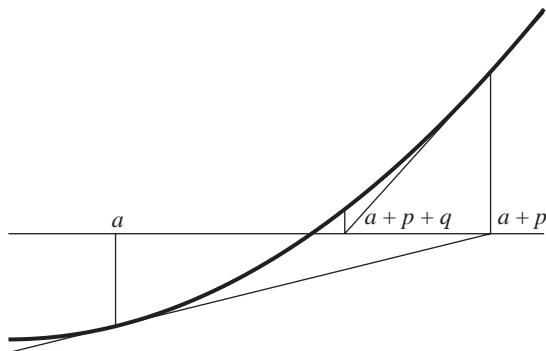


Figure 3 Newton's method.

2.3.1 Tangents and Convergence

Newton's method has a geometrical interpretation, which Newton himself did not give, in terms of the graph of a function f . A root x of the equation $f(x) = 0$ corresponds to a point where the curve with equation $y = f(x)$ intersects the x -axis. If you start with an approximate value a for x and set $p = x - a$, as we did above, then when you substitute $a + p$ for x to obtain a new function $g(p)$, you are effectively moving the origin from $(0, 0)$ to the point $(a, 0)$. Then when you forget all powers of p other than the constant and linear terms, you are finding the best linear approximation to the function g —which, geometrically speaking, is the tangent line to g at the point $(0, g(0))$. Thus, the approximate value you obtain for p is the x -coordinate of the point where the tangent at $(0, g(0))$ crosses the x -axis. Adding a to this value returns the origin to $(0, 0)$ and gives the new approximation to the root of f . This is why Newton's method is often called the *tangent method* (figure 3). And one can now see that the new approximation will definitely be better than the old one if the tangent to f at $(a, f(a))$ intersects the x -axis at a point that lies between a and the point where the curve $y = f(x)$ intersects the x -axis.

As it happens, this is not the case for Newton's choice of the value $a = 2$ above, but it is true for the approximate value 2.1 and for all subsequent ones. Geometrically, the favorable situation occurs if the point $(a, f(a))$ lies above the x -axis in a convex part of the curve that crosses the x -axis or below the x -axis in a concave part of the curve that crosses the x -axis. Under these circumstances, and provided the root is not a multiple one, the convergence is *quadratic*, meaning that the error at each stage is roughly the square of the error at the previous stage—or, equivalently, the

approximation is valid to a number of decimal places that roughly doubles at each stage. This is enormously fast.

The choice of the initial approximation value is obviously important, and raises unexpectedly subtle questions. These are clearer if we look at *complex* polynomials and their complex roots. Newton's method can be easily adapted to this more general context. Suppose that z is a root of some complex polynomial and that z_0 is an initial approximation for z . Newton's method then gives us a sequence z_0, z_1, z_2, \dots , which may or may not converge to z . We define the *domain of attraction*, denoted $A(z)$, to be the set of all complex numbers z_0 such that the resulting sequence does indeed converge to z . How do we determine $A(z)$?

The first person to ask this problem was CAYLEY [VI.46], in 1879. He noticed that the solution is easy for quadratic polynomials but difficult as soon as the degree is 3 or more. For example, the domains of attraction of the roots ± 1 of the polynomial $z^2 - 1$ are the open half-planes bounded by the vertical axis, but the domains corresponding to the roots 1, ω , and ω^2 of $z^3 - 1$ are extremely complicated sets. They were described by Julia in 1918—such subsets are now called *fractal sets*. Newton's method and fractal sets are discussed further in DYNAMICS [IV.15].

2.3.2 Recurrence Formulas

At each stage of his method, Newton had to produce a new equation, but in 1690 Raphson noticed that this was not really necessary. For particular examples, he gave single formulas that could be used at each step, but his basic observation applies in general and leads to a general formula for every case, which one can easily obtain using the interpretation in terms of tangents. Indeed, the tangent to the curve $y = f(x)$ at the point of x -coordinate a has the equation $y - f(a) = f'(a)(x - a)$, and it cuts the x -axis at the point with x -coordinate $a - f(a)/f'(a)$. What we now call the *Newton-Raphson method* springs from this simple formula. One starts with an initial approximation $a_0 = a$ and then defines successive approximations using the recurrence formula

$$a_{n+1} = a_n - \frac{f(a_n)}{f'(a_n)}.$$

As an example, let us consider the function $f(x) = x^2 - c$. Here, Newton's method provides a sequence of approximations of the square root \sqrt{c} of c , given by the recurrence formula $a_{n+1} = \frac{1}{2}(a_n + c/a_n)$ (which

we obtain by substituting $x^2 + c$ for f in the general formula above). This method for approximating square roots was known by Heron of Alexandria in the first century. Note that if a_0 is close to \sqrt{c} , then c/a_0 is also close, \sqrt{c} lies between them, and $a_1 = \frac{1}{2}(a_0 + c/a_0)$ is their arithmetic mean.

3 Does an Algorithm Always Exist?

3.1 Hilbert's Tenth Problem: The Need for Formalization

In 1900, at the Second International Congress of Mathematicians, HILBERT [VI.63] proposed a list of twenty-three problems. These problems, and Hilbert's works in general, had a huge influence on mathematics during the twentieth century (Gray 2000). We are interested here in *Hilbert's tenth problem*: given a Diophantine equation, that is, a polynomial equation with any number of indeterminates and with integer coefficients, "a process is sought by which it can be determined, in a finite number of operations, whether the equation is solvable in integral numbers." In other words, we have to find an algorithm which tells us, for any Diophantine equation, whether or not it has at least one integer solution. Of course, for many Diophantine equations it is easy to find solutions, or to prove that no solutions exist. However, this is by no means always the case: consider, for instance, the Fermat equation $x^n + y^n = z^n$ ($n \geq 3$). (Even before the solution of FERMAT'S LAST THEOREM [V.12] an algorithm was known for determining for any specific n whether this equation had a solution. However, one could not call it easy.)

If Hilbert's tenth problem has a positive answer, then one can demonstrate it by exhibiting a "process" of the sort that Hilbert asked for. To do this, it is not necessary to have a precise understanding of what a "process" is. However, if you want to give a *negative* answer, then you have to show that *no algorithm exists*, and for that you need to say precisely what counts as an algorithm. In section 1.2 we gave a definition that seems to be reasonably precise, but it is not precise enough to enable us to think about Hilbert's tenth problem. What kind of rules are we allowed to use in an algorithm? How can we be sure that no algorithm achieves a certain task, rather than just that we are unable to find one?

3.2 Recursive Functions: Church's Thesis

What we need is a *formal* definition of the notion of an algorithm. In the seventeenth century, LEIBNIZ [VI.15]

envisaged a universal language that would allow one to reduce mathematical proofs to simple computations. Then, during the nineteenth century, logicians such as Charles Babbage, BOOLE [VI.43], FREGE [VI.56], and PEANO [VI.62] tried to formalize mathematical reasoning by an "algebraization" of logic. Finally, between 1931 and 1936, GÖDEL [VI.92], CHURCH [VI.89], and Stephen Kleene introduced the notion of *recursive functions* (see Davis (1965), which contains the original texts). Roughly speaking, a recursive function is one that can be calculated by means of an algorithm, but the *definition* of recursive functions is different, and is completely precise.

3.2.1 Primitive Recursive Functions

Another rough definition of a recursive function is as follows: a recursive function is one that has an inductive definition. To give an idea of what this means, let us consider addition and multiplication as functions from $\mathbb{N} \times \mathbb{N}$ to \mathbb{N} . To emphasize this, we shall write $\text{sum}(x, y)$ and $\text{prod}(x, y)$ for $x + y$ and xy , respectively.

A familiar fact about multiplication is that it is "repeated addition." Let us examine this more precisely. We can define the function "prod" in terms of the function "sum" by means of the following two rules: $\text{prod}(1, y) = y$ and $\text{prod}(x + 1, y) = \text{sum}(\text{prod}(x, y), y)$. Thus, if you know $\text{prod}(x, y)$ and you know how to calculate sums, then you know $\text{prod}(x + 1, y)$. Since you also know the "base case" $\text{prod}(1, y)$, a simple inductive argument shows that these simple rules completely determine the function "prod."

We have just seen how one function can be "recursively defined" in terms of another. We now want to understand the class of *all* functions from \mathbb{N}^n to \mathbb{N} that can be built up in a few basic ways, of which recursion is the most important. We shall refer to functions from \mathbb{N}^n to \mathbb{N} as *n-ary functions*.

To begin with, we need an initial stock of functions out of which the rest will be built. It turns out that a very simple set of functions is enough. Most basic are the *constant functions*: that is, functions that take every n -tuple in \mathbb{N}^n to some fixed positive integer c . Another very simple function, but the function that allows us to create much more interesting ones, is the *successor function*, which takes a positive integer n to the next one, $n + 1$. Finally, we have *projection functions*: the function U_k^n takes a sequence (x_1, \dots, x_n) in \mathbb{N}^n and maps it to the k th coordinate x_k .

We then have two ways of constructing functions from other functions. The first is *substitution*. Given an m -ary function ϕ and m n -ary functions ψ_1, \dots, ψ_m , one defines an n -ary function by $(x_1, \dots, x_n) \mapsto \phi(\psi_1(x_1, \dots, x_n), \dots, \psi_m(x_1, \dots, x_n))$. For example, $(x + y)^2 = \text{prod}(\text{sum}(x, y), \text{sum}(x, y))$, so we can obtain the function $(x, y) \mapsto (x + y)^2$ from the functions “sum” and “prod” by means of substitution.

The second method of construction is called *primitive recursion*. This is a more general form of the inductive method we used above in order to construct the function “prod” from the function “sum.” Given an $(n - 1)$ -ary function ψ and an $(n + 1)$ -ary function μ , one defines an n -ary function ϕ by saying that $\phi(1, x_2, \dots, x_n) = \psi(x_2, \dots, x_n)$ and $\phi(k + 1, x_2, \dots, x_n) = \mu(k, \phi(k, x_2, \dots, x_n), x_2, \dots, x_n)$. In other words, ψ tells you the “initial values” of ϕ (the values when the first coordinate is 1) and μ tells you how to work out $\phi(k + 1, x_2, \dots, x_n)$ in terms of $\phi(k, x_2, \dots, x_n)$, x_2, \dots, x_n and k . (The sum-product example was simpler because we did not have a dependence on k .)

A *primitive recursive function* is any function that can be built from the initial stock of functions using the two operations, substitution and primitive recursion, that we have just described.

3.2.2 Recursive Functions

If you think for a while about primitive recursion and know a small amount about programming computers, you should be able to convince yourself that they are *effectively computable*: that is, that for any primitive recursive function there is an algorithm for computing it. (For example, the operation of primitive recursion can usually be realized in a rather direct way as a FOR loop.)

How about the converse? Are all computable functions primitive recursive? Consider, for example, the function that takes the positive integer n to p_n , the n th prime number. It is not hard to devise a simple algorithm for computing p_n , and it is then a good exercise (if you want to understand primitive recursion) to convert this algorithm into a proof that the function is primitive recursive.

However, it turns out that this function is not typical: there are computable functions that are not primitive recursive. In 1928, Wilhelm Ackermann defined a function, now known as the *Ackermann function*, that has a “doubly inductive” definition. The following function is

not quite the same as Ackermann’s, but it is very similar. It is the function $A(x, y)$ that is determined by the following recurrence rules:

- (i) $A(1, y) = y + 2$ for every y ;
- (ii) $A(x, 1) = 2$ for every x ;
- (iii) $A(x + 1, y + 1) = A(x, A(x + 1, y))$ whenever $x > 1$ and $y > 1$.

For example, $A(2, y + 1) = A(1, A(2, y)) = A(2, y) + 2$. From this and the fact that $A(2, 1) = 2$ it follows that $A(2, y) = 2y$ for every y . In a similar way one can show that $A(3, y) = 2^y$, and in general that for each x the function that takes y to $A(x + 1, y)$ “iterates” the function that takes y to $A(x, y)$. This means that the values of $A(x, y)$ are extremely large even when x and y are fairly small. For example, $A(4, y + 1) = 2^{A(4, y)}$, so in general $A(4, y)$ is given by an “exponential tower” of height y . We have $A(4, 1) = 2$, $A(4, 2) = 2^2 = 4$, $A(4, 3) = 2^4 = 16$, $A(4, 4) = 2^{16} = 65\,536$, and $A(4, 5) = 2^{65\,536}$, which is too large a number for its decimal notation to be reproduced here.

It can be shown that for every primitive recursive function ϕ there is some x such that the function $A(x, y)$ grows faster than $\phi(y)$. This is proved by an inductive argument. To oversimplify slightly, if $\psi(y)$ and $\mu(y)$ have already been shown to grow more slowly than $A(x, y)$, then one can show that the function ϕ produced from them by primitive recursion also grows more slowly. This allows us to define a “diagonal” function $A(y) = A(y, y)$ that is not primitive recursive because it grows faster than any of the functions $A(x, y)$.

If we are trying to understand in a precise way which functions can be calculated algorithmically, then our definition will surely have to encompass functions like the Ackermann function, since they can in principle be computed. Therefore, we must consider a larger class of functions than just the primitive recursive ones. This is what Gödel, Church, and Kleene did, and they obtained in different ways the same class of *recursive functions*. For instance, Kleene added a third method of construction, which he called *minimization*. If f is an $(n + 1)$ -ary function, one defines an n -ary function g by taking $g(x_1, \dots, x_n)$ to be the smallest y such that $f(x_1, \dots, x_n, y) = 0$. (If there is no such y , one regards g as undefined for (x_1, \dots, x_n) . We shall ignore this complication in what follows.)

It turns out that, not only is the Ackermann function recursive, but so are all functions that one can write

a computer program to calculate. So this gives us the formal definition of computability that we did not have before.

3.2.3 Effective Calculability

With such a class of recursive functions, Church claimed that the class of “effectively calculable” functions is exactly the class of recursive functions. Church’s thesis is widely believed, but this is a conviction that cannot be proved since the notion of recursive function is a mathematically precise concept while that of an effectively calculable function is an intuitive notion, actually quite like that of “algorithm.” Church’s statement lies in the realm of metamathematics and is now called *Church’s thesis*.

3.3 Turing Machines

One of the strongest pieces of evidence for Church’s thesis is that in 1936 TURING [VI.94] found a very different-looking way of formalizing the notion of an algorithm, which he showed was equivalent. That is, every function that was computable in his new sense was recursive and vice versa. His approach was to define a notion that is now called a *Turing machine*, which can be thought of as an extremely primitive computer, and which played an important part in the development of actual computers. Indeed, functions that are computable by Turing machines are precisely those that can be programmed on a computer. The primitive architecture of Turing machines does not make them any less powerful: it merely means that in practice they would be too cumbersome to program or to implement in hardware. Since recursive functions are the same as Turing-computable functions, it follows that recursive functions too are those functions that can be programmed on a computer, so to disbelieve Church’s thesis would be to maintain that there are some “effective procedures” that cannot be converted into computer programs—which seems rather implausible. A description of Turing machines can be found in COMPUTATIONAL COMPLEXITY [IV.21 §1].

Turing introduced his machines in response to a question that generalized Hilbert’s tenth problem. The *Entscheidungsproblem*, or *decision problem*, was also asked by Hilbert, in 1922. He wanted to know whether there was a “mechanical process” by which one could determine whether any given mathematical statement could be proved. In order to think about this, Turing

needed a precise notion of what constituted a “mechanical process.” Once he had defined Turing machines, he was able to show by means of a fairly straightforward diagonal argument that the answer to Hilbert’s question was no. His argument is outlined in THE INSOLUBILITY OF THE HALTING PROBLEM [V.23].

4 Properties of Algorithms

4.1 Iteration versus Recursion

As previously mentioned, we often encounter computation rules which define each element of a sequence in terms of the preceding elements. This gives rise to two different ways of carrying out the computation. The first is *iteration*: one computes the first terms, then one obtains succeeding terms by means of a recurrence formula. The second is *recursion*, a procedure which seems circular at first because one defines a procedure in terms of itself. However, this is allowed because the procedure calls on itself with smaller values of the variables. The concept of recursion is subtle and powerful. Let us try to clarify the difference between recursion and iteration with some examples.

Suppose that we wish to compute $n! = 1 \cdot 2 \cdot 3 \cdot \dots \cdot (n-1) \cdot n$. An obvious way of doing it is to note the recurrence relation $n! = n \cdot (n-1)!$ and the initial value $1! = 1$. Having done so, one could then compute successively the numbers $2!$, $3!$, $4!$, and so on until one reached $n!$, which would be the iterative approach. Alternatively, one could say that if $\text{fact}(n)$ is the result of a procedure that leads to $n!$, then $\text{fact}(n) = n \times \text{fact}(n-1)$, which would be a recursive procedure. The second approach says that to obtain $n!$ it suffices to know how to obtain $(n-1)!$, and to obtain $(n-1)!$ it suffices to know how to obtain $(n-2)!$, and so on. Since one knows that $1! = 1$, one can obtain $n!$. Thus, recursion is a bit like iteration but thought of “backwards.”

In some ways this example is too simple to show clearly the difference between the two procedures. Moreover, if one wishes to compute $n!$, then iteration seems simpler and more natural than recursion. We now look at an example where recursion is far simpler than iteration.

4.1.1 The Tower of Hanoi

The Tower of Hanoi is a problem that goes back to Édouard Lucas in 1884. One is given n disks, all of different sizes and each with a hole in the middle, stacked on a peg A in order of size, with the largest one at the

bottom. We also have two empty pegs B and C. The problem is to move the stack from peg A to peg B while obeying the following rules. One is allowed to move just one disk at a time, and each move consists in taking the top disk from one of the pegs and putting it onto another peg. In addition, no disk may ever be placed above a smaller disk.

The problem is easy if you have just three disks, but becomes rapidly harder as the number of disks increases. However, with the help of recursion one can see very quickly that an algorithm exists for moving the disks in the required way. Indeed, suppose that we know a procedure $H(n-1)$ that solves the problem for $n-1$ disks. Here is a procedure $H(n)$ for n disks: move the first $n-1$ disks on top of A to C with the procedure $H(n-1)$, then move the last disk on A to B, and finally apply once more the procedure $H(n-1)$ to move all the disks from C to B. If we write $H_{AB}(n)$ for the procedure that moves n disks from peg A to peg B according to the rules, then we can represent this recursion symbolically as

$$H_{AB}(n) = H_{AC}(n-1)H_{AB}(1)H_{BC}(n-1).$$

Thus, $H_{AB}(n)$ is deduced from $H_{AC}(n-1)$ and $H_{BC}(n-1)$, which are clearly equivalent to $H_{AB}(n-1)$. Since $H_{AB}(1)$ is certainly easy, we have the full recursion.

One can easily check by induction that this procedure takes $2^n - 1$ moves—moreover, it turns out that the task cannot be accomplished in fewer moves. Thus, the number of moves is an exponential function of n , so for large n the procedure will be very long.

Furthermore, the larger n is, the more memory one must use to keep track of where one is in the procedure. By contrast, if we wish to carry out an iteration during an iterative procedure, it is usually enough to know just the result of the previous iteration. Thus, the most we need to remember is the result of one iteration. There is in fact an iterative procedure for the Tower of Hanoi as well. It is easy to describe, but it is much less obvious that it actually solves the problem. It encodes the positions of the n disks as an n -bit sequence and at each step applies a very simple rule to obtain the next n -bit sequence. This rule makes no reference to how many steps have so far taken place, and therefore the amount of memory needed, beyond that required to store the positions of the disks, is very small.

4.1.2 The Extended Euclid Algorithm

Euclid's algorithm is another example that lends itself in a very natural way to a recursive procedure. Recall

that if a and b are two positive integers, then we can write $a = qb + r$ with $0 \leq r < b$. The algorithm depended on the observation that $\gcd(a, b) = \gcd(b, r)$. Since the remainder r can be calculated easily from a and b , and since the pair (b, r) is smaller than the pair (a, b) , this gives us a recursive procedure, which stops when we reach a pair of the form $(a, 0)$.

An important extension of Euclid's algorithm is *Bézout's lemma*, which states that for any pair of positive integers (a, b) there exist (not necessarily positive) integers u and v such that

$$ua + vb = d = \gcd(a, b).$$

How can we obtain such integers u and v ? The answer is given by the *extended Euclid algorithm*, which again can be defined using recursion. Suppose we can find a pair (u', v') that works for b and r : that is, $u'b + v'r = d$. Since $a = qb + r$, we can substitute $r = a - qb$ into this equation and deduce that $d = u'b + v'(a - qb) = v'a + (u' - v'q)b$. Thus, setting $u = v'$ and $v = u' - v'q$, we have $ua + vb = d$. Since a pair (u, v) that works for a and b can be easily calculated from a pair (u', v') that works for the smaller b and r , this gives us a recursive procedure. The “bottom” of the recursion is when $r = 0$, in which case we know that $1b + 0r = d$. Once we reach this, we can “run back up” through Euclid's algorithm, successively modifying our pair (u, v) according to the rule just given. Notice, incidentally, that the fact that this procedure exists is a proof of Bézout's lemma.

4.2 Complexity

So far we have considered algorithms in a theoretical way and ignored their obvious practical importance. However, the mere existence of an algorithm for carrying out a certain task does not guarantee that your computer can do it, because some algorithms take so many steps that no computer can implement them (unless you are prepared to wait billions of years for the answer). The *complexity* of an algorithm is, loosely speaking, the number of steps it takes to complete its task (as a function of the size of the input). More precisely, this is the *time complexity* of the algorithm. There is also its *space complexity*, which measures the maximum amount of memory a computer needs in order to implement it. *Complexity theory* is the study of the computational resources that are needed to carry out various tasks. It is discussed in detail in COMPUTATIONAL COMPLEXITY [IV.21]—here we shall give a hint of it by examining the complexity of one algorithm.

4.2.1 The Complexity of Euclid's Algorithm

The length of time that a computer will take to implement Euclid's algorithm is closely related to the number of times one needs to compute quotients and remainders: that is, to the number of times that the recursive procedure calls on itself. Of course, this number depends in turn on the size of the numbers a and b whose gcd is to be determined. An initial observation is that if $0 < b \leq a$, then the remainder in the division of a by b is less than $a/2$. To see this, notice that if $b \geq a/2$ then the remainder is $a - b$, which is at most $a/2$, whereas if $b \leq a/2$ then we know that the remainder is at most b and so is again at most $a/2$. It follows that after two steps of calculating the remainder, one arrives at a pair where the larger number is at most half what it was before. From this it is easy to show that the number of such calculations needed is at most $2 \log_2 a + 1$, which is roughly proportional to the number of digits of a . Since this number is far smaller than a itself, the algorithm can be used easily for very large numbers, which gives it great practical utility to go with its theoretical significance.

The number of divisions needed in the worst case does not appear to have been studied until the first half of the nineteenth century: the above bound of $2 \log_2 a + 1$ was given by Pierre-Joseph-Étienne Finck in 1841. It is in fact not hard to improve this result slightly and prove that the algorithm takes longest when a and b are consecutive Fibonacci numbers. This implies that the number of divisions needed is never more than $\log_\phi a + 1$, where ϕ is the golden ratio.

Euclid's algorithm also has low space complexity: once one has replaced a pair (a, b) by a new pair (b, r) , one can forget the original pair, so at any stage one does not have to hold very much in one's memory (or store it in the memory of one's computer). By contrast, the extended Euclid algorithm appears to require one to remember the entire sequence of calculations that leads to the gcd d of a and b , so that one can make a series of substitutions and eventually find u and v such that $ua + vb = d$. However, a closer look at it reveals that one can perform it while keeping track of only a few numbers at any one time.

Let us see how this works with an example. We shall set $a = 38$, $b = 21$, and find integers u and v such that $38u + 21v = 1$. We begin by writing down the first step of Euclid's algorithm:

$$38 = 1 \times 21 + 17.$$

This tells us that $17 = 38 - 21$. Now we write down the second step:

$$21 = 1 \times 17 + 4.$$

We know how to write 17 in terms of 38 and 21, so let us do a substitution:

$$21 = 1 \times (38 - 21) + 4.$$

Rearranging this, we discover that $4 = 2 \times 21 - 38$. Now we write down the third step of Euclid's algorithm:

$$17 = 4 \times 4 + 1.$$

We know how to write 17 and 4 in terms of 38 and 21, so let us substitute again:

$$38 - 21 = 4 \times (2 \times 21 - 38) + 1.$$

Rearranging this, we find that $1 = 5 \times 38 - 9 \times 21$, and we have finished.

If you think about this procedure, you will see that at each stage one just has to keep track of how two numbers are expressed in terms of a and b . Thus, the space complexity of the extended Euclid algorithm is small if you implement it properly.

5 Modern Aspects of Algorithms

5.1 Algorithms and Chance

Earlier it was remarked that the notion of algorithm has continued to develop even since its formalization in the 1920s and 1930s. One of the main reasons for this has been the realization that *randomness* can be a very useful tool in algorithms. This may seem puzzling at first, since algorithms as we have described them are deterministic procedures; in a moment we shall give an example that illustrates how randomness can be used. A second reason is the recent development of the notion of a *quantum algorithm*: for more about this, see QUANTUM COMPUTATION [III.76].

The following simple example illustrates how chance can be useful. Given an integer n , we shall define a function $f(n)$ that is not too hard to calculate but is difficult to analyze. If n has d digits, then you approximate \sqrt{n} to the point where the first d digits after the decimal point are correct (using Newton's method, say), and let $f(n)$ equal the d th digit. Now suppose that you wish to know roughly what proportion of numbers n between 10^{30} and 10^{31} have $f(n) = 0$. There does not seem to be a good way of determining this theoretically, but calculating it on a computer looks very hard, too, as there are so many numbers between 10^{30} and 10^{31} . However, if one chooses a random sample of 10000

numbers between 10^{30} and 10^{31} and does the calculation for just those numbers, then with high probability the proportion of those numbers with $f(n) = 0$ will be roughly the same as the proportion of all numbers in the range with $f(n) = 0$. Thus, if you do not demand absolute certainty but instead are satisfied with a very small error probability, then you can achieve your goal with much more modest computational resources.

5.1.1 Pseudorandom Numbers

How, though, does one use a deterministic computer to select ten thousand random numbers between 10^{30} and 10^{31} ? The answer is that one does not in fact need to: it is almost always good enough to make a *pseudorandom* selection instead. The basic idea is well-illustrated by a method proposed by VON NEUMANN [VI.91] in the mid 1940s. One begins with a $2n$ -digit integer a , called the “seed,” calculates a^2 , and extracts from a^2 a new $2n$ -digit number b by taking all the digits of a^2 from the $(n + 1)$ st to the $3n$ th. One then repeats the procedure for b , and so on. Because of the way multiplication jumbles up digits, the resulting sequence of $2n$ -digit numbers is very hard to distinguish from a truly random sequence, and can be used in randomized algorithms.

There are many other ways of producing pseudorandom sequences, and this raises an obvious question: what properties should a sequence have for us to regard it as pseudorandom? This turns out to be a complex question, and several different answers have been proposed. Randomized algorithms and pseudorandomness are discussed in depth in COMPUTATIONAL COMPLEXITY [IV.21 §§6, 7], and a formal definition of “pseudorandom generators” can be found there. (See also COMPUTATIONAL NUMBER THEORY [IV.5 §2] for an account of a famous randomized algorithm for testing whether a number is prime.) Here, let us discuss a similar question about infinite sequences of zeros and ones. When should we regard such a sequence as “random”?

Again, many different answers have been suggested. One idea is to consider simple statistical tests: we would expect that in the long run the frequency of zeros should be roughly the same as that of ones, and more generally that any small subsequence such as 00110 should appear with the “right” frequency (which for this sequence would be $\frac{1}{32}$ since it has length 5).

It is perfectly possible, however, for a sequence to pass these simple tests but to be generated by a deterministic procedure. If one is trying to decide whether a sequence of zeros and ones is *actually* random—that is, produced by some means such as tossing a

coin—then we will be very suspicious of a sequence if we can identify an algorithm that produces the same sequence. For example, we would reject a sequence that was derived in a simple way from the digits of π , even if it passed the statistical tests. However, merely to ask that a sequence cannot be produced by a recursive procedure does not give a good test for randomness: for example, if one takes any such sequence and alternates the terms of that sequence with zeros, one then obtains a new sequence that is far from random, but which still cannot be produced recursively.

For this reason, von Mises suggested in 1919 that a sequence of zeros and ones should be called random if it is not only the case that the limit of the frequency of ones is $\frac{1}{2}$, but also that the same is true for any subsequence that can be extracted “by means of a reasonable procedure.” In 1940 Church made this more precise by translating “by means of a reasonable procedure” into “by means of a recursive function.” However, even this condition is too weak: there are such sequences that do not satisfy the “law of the iterated logarithm” (something that a random sequence would satisfy). Currently, the so-called Martin-Löf thesis, formulated in 1966, is one of the most commonly used definitions of randomness: a random sequence is a sequence that satisfies all the “effective statistical sequential tests,” a notion that we cannot formulate precisely here, but which uses in an essential manner the notion of recursive function. By contrast with Church’s thesis, with which almost every mathematician agrees, the Martin-Löf thesis is still very much under discussion.

5.2 The Influence of Algorithms on Contemporary Mathematics

Throughout its history, mathematics has concerned itself with problems of existence. For example, are there TRANSCENDENTAL NUMBERS [III.43], that is, numbers that are not the root of any polynomial with integer coefficients? There are two kinds of answers to such questions: either one actually exhibits a number such as π and proves that it is transcendental (this was done by Carl Lindemann in 1873), or one gives an “indirect existence proof,” such as CANTOR’s [VI.54] demonstration that there are “far more” real numbers than there are roots of polynomials with integer coefficients (see COUNTABLE AND UNCOUNTABLE SETS [III.11]), which shows in particular that some real numbers must be transcendental.

5.2.1 Constructivist Schools

In around 1910, under the influence of BROUWER [VI.75], the INTUITIONIST SCHOOL [II.7 §3.1] of mathematics arose, which rejected the principle of the excluded middle, which is the principle that every mathematical assertion is either true or false. In particular, Brouwer did not accept that the existence of a mathematical object such as a transcendental number is proved by the fact that its nonexistence would lead to a contradiction. This was the first of several “constructivist” schools, for which an object exists if and only if it can be constructed explicitly.

Not many working mathematicians subscribe to these principles, but almost all would agree that there is an important difference between constructive proofs and indirect proofs of existence, a difference that has come to seem more important with the rise of computer science. This has added a further level of refinement: sometimes, even if you know that a mathematical object can be produced algorithmically, you still care whether the algorithm can be made to work in a reasonably short time.

5.2.2 Effective Results

In number theory there is an important distinction between “effective” and “ineffective” results. For example, MORDELL’S CONJECTURE [V.31], proposed in 1922 and finally proved by Faltings in 1983, states that a smooth rational plane curve of degree $n > 3$ has at most finitely many points with rational coefficients. Among its many consequences is that the Fermat equation $x^n + y^n = z^n$ has only finitely many integral solutions for each $n \geq 4$. (Of course, we now know that it has no nontrivial solutions, but the Mordell conjecture was proved before Fermat’s last theorem, and it has many other consequences.) However, Faltings’s proof is *ineffective*, which means that it does not give any information about how many solutions there are (except that there are not infinitely many), or how large they can be, so one cannot use a computer to find them all and know that one has finished the job. There are many other very important proofs in number theory that are ineffective, and replacing any one of them with an effective argument would be a major breakthrough.

A completely different set of issues was raised by another solution to a famous open problem, the FOUR-COLOR THEOREM [V.14], which was conjectured by Francis Guthrie, a student of DE MORGAN [VI.38], in 1852 and proved in 1976 by Appel and Haken, with a proof

that made essential use of computers. They began with a theoretical argument that reduced the problem to checking finitely many cases, but the number of cases was so large that it could not be done by hand and was instead done by computers. But how should we judge such a proof? Can we be sure that the computer has been programmed correctly? And even if it has, how do we know with a computation of that size that the computer has operated correctly? And does a proof that relies on a computer really tell us *why* the theorem is true? These questions continue to be debated today.

Further Reading

- Archimedes. 2002. *The Works of Archimedes*, translated by T. L. Heath. London: Dover. Originally published 1897, Cambridge University Press, Cambridge.
- Chabert, J.-L., ed. 1999. *A History of Algorithms: From the Pebble to the Microchip*. Berlin: Springer
- Davis, M., ed. 1965. *The Undecidable*. New York: The Raven Press.
- Euclid. 1956. *The Thirteen Books of Euclid’s Elements*, translated by T. L. Heath (3 vols.), 2nd edn. London: Dover. Originally published 1929, Cambridge University Press, Cambridge.
- Gray, J. J. 2000. *The Hilbert Challenge*. Oxford: Oxford University Press.
- Newton, I. 1969. *The Mathematical Papers of Isaac Newton*, edited by D. T. Whiteside, volume 3 (1670–73), pp. 43–47. Cambridge: Cambridge University Press.

II.5 The Development of Rigor in Mathematical Analysis

Tom Archibald

1 Background

This article is about how rigor was introduced into mathematical analysis. The question is a complicated one, since mathematical practice has changed considerably, especially in the period between the founding of the calculus (shortly before 1700) and the early twentieth century. In a sense, the basic criteria for what constitutes a correct and logical argument have not altered, but the circumstances under which one would require such an argument, and even to some degree the purpose of the argument, have altered with time. The voluminous and successful mathematical analysis of the 1700s, associated with names such as Johann and Daniel BERNOULLI [VI.18], EULER [VI.19], and LAGRANGE [VI.22], lacked foundational clarity in ways that were criticized and remedied in subsequent periods. By

around 1910 a general consensus had emerged about how to make arguments in analysis rigorous.

Mathematics consists of more than techniques for calculation, methods for describing important features of geometric objects, and models of worldly phenomena. Almost all working mathematicians today are trained in, and concerned with, the production of rigorous arguments that justify their conclusions. These conclusions are usually framed as *theorems*, which are statements of fact, accompanied by an argument, or proof, that the theorem is indeed true. Here is a simple example: every positive whole number that is divisible by 6 is also divisible by 2. Running through the six times table (6, 12, 18, 24, ...) we see that each number is even, which makes the statement easy enough to believe. A possible justification of it would be to say that since 6 is divisible by 2, then every number divisible by 6 must also be divisible by 2.

Such a justification might or might not be thought of as a thorough proof, depending on the reader. For on hearing the justification we can raise questions: is it always true that if a , b , and c are three positive whole numbers such that c is divisible by b and b is divisible by a , then c is divisible by a ? What is divisibility exactly? What is a whole number? The mathematician deals with such questions by precisely defining concepts (such as divisibility of one number by another), basing the definitions on a smallish number of undefined terms ("whole number" might be one, though it is possible to start even further back, with sets). For example, one could define a number n to be divisible by a number m if and only if there exists an integer q such that $qm = n$. Using this definition, we can give a more precise proof: if n is divisible by 6, then $n = 6q$ for some q , and therefore $n = 2(3q)$, which proves that n is divisible by 2. Thus we have used the definitions to show that the definition of divisibility by 2 holds whenever the definition of divisibility by 6 holds.

Historically, mathematical writers have been satisfied with varying levels of rigor. Results and methods have often been widely used without a full justification of the kind just outlined, particularly in bodies of mathematical thought that are new and rapidly developing. Some ancient cultures, the Egyptians for example, had methods for multiplication and division, but no justification of these methods has survived and it does not seem especially likely that formal justification existed. The methods were probably accepted simply because they worked, rather than because there was a thorough argument justifying them.

By the middle of the seventeenth century, European mathematical writers who were engaged in research were well-acquainted with the model of rigorous mathematical argument supplied by EUCLID's [VI.2] *Elements*. The kind of deductive, or synthetic, argument we illustrated earlier would have been described as a proof *more geometrico*—in the geometrical way. While Euclid's arguments, assumptions, and definitions are not wholly rigorous by today's standards, the basic idea was clear: one proceeds from clear definitions and generally agreed basic ideas (such as that the whole is greater than the part) to deduce theorems (also called propositions) in a step-by-step manner, not bringing in anything extra (either on the sly or unintentionally). This classical model of geometric argument was widely used in reasoning about whole numbers (for example by FERMAT [VI.12]), in analytic geometry (DESCARTES [VI.11]), and in mechanics (Galileo).

This article is about rigor in *analysis*, a term which itself has had a shifting meaning. Coming from ancient origins, by around 1600 the term was used to refer to mathematics in which one worked with an unknown (something we would now write as x) to do a calculation or find a length. In other words, it was closely related to algebra, though the notion was imported into geometry by Descartes and others. However, over the course of the eighteenth century the word came to be associated with the calculus, which was the principal area of application of analytic techniques. When we talk about rigor in analysis it is the rigorous theory of the mathematics associated with differential and integral calculus that we are principally discussing. In the third quarter of the seventeenth century rival methods for the differential and integral calculus were devised by NEWTON [VI.14] and LEIBNIZ [VI.15], who thereby synthesized and extended a considerable amount of earlier work concerned with tangents and normals to curves and with the areas of regions bounded by curves. The techniques were highly successful, and were extended readily in a variety of directions, most notably in mechanics and in differential equations.

The key common feature of this research was the use of infinities: in some sense, it involved devising methods for combining infinitely many infinitely small quantities to get a finite answer. For example, suppose we divide the circumference of a circle into a (large) number of equal parts by marking off points at equal distances, then joining the points and creating triangles by joining the points to the center. Adding up the areas of the triangles approximates the circular area, and the

more points we use the better the approximation. If we imagine infinitely many of these inscribed triangles, the area of each will be “infinitely small” or *infinitesimal*. But because the total involves adding up infinitely many of them, it may be that we get a finite positive total (rather than just 0, from adding up infinitely many zeros, or an infinite number, as we would get if we added the same finite number to itself infinitely many times). Many techniques for doing such calculations were devised, though the interpretation of what was taking place varied. Were the infinities involved “real” or merely “potential”? If something is “really” infinitesimal, is it just zero? Aristotelian writers had abhorred actual infinities, and complaints about them were common at the time.

Newton, Leibniz, and their immediate followers provided mathematical arguments to justify these methods. However, the introduction of techniques involving reasoning with infinitely small objects, limiting processes, infinite sums, and so forth meant that the founders of the calculus were exploring new ground in their arguments, and the comprehensibility of these arguments was frequently compromised by vague terms, or the drawing of one conclusion when another might seem to follow equally well. The objects they were discussing included infinitesimals (quantities infinitely smaller than those we experience directly), ratios of vanishingly small quantities (i.e., fractions in, or approaching, the form $0/0$), and finite sums of infinitely many positive terms. Taylor series representations, in particular, provoked a variety of questions. A function may be written as a series in such a way that the series, when viewed as a function, will have, at a given point $x = a$, the same value as the function, the same rate of change (or first derivative), and the same higher-order derivatives to arbitrary order:

$$f(x) = f(a) + f'(a)(x - a) + \frac{1}{2}f''(a)(x - a)^2 + \cdots$$

For example, $\sin x = x - x^3/3! + x^5/5! + \cdots$, a fact already known to Newton though such series are now named after Newton’s disciple BROOK TAYLOR [VI.16].

One problem with early arguments was that the terms being discussed were used in different ways by different writers. Other problems arose from this lack of clarity, since it concealed a variety of issues. Perhaps the most important of these was that an argument could fail to work in one context, even though a very similar argument worked perfectly well in another. In time, this led to serious problems in extending analysis. Eventually, analysis became fully rigorous and these

difficulties were solved, but the process was a long one and it was complete only by the beginning of the twentieth century.

Let us consider some examples of the kinds of difficulties that arose from the very beginning, using a result of Leibniz. Suppose we have two variables, u and v , each of which changes when another variable, x , changes. An infinitesimal change in x is denoted dx , the differential of x . The differential is an infinitesimal quantity, thought of as a geometrical magnitude, such as a length, for example. This was imagined to be combined or compared with other magnitudes in the usual ways (two lengths can be added, have a ratio, and so on). When x changes to $x + dx$, u and v change to $u + du$ and $v + dv$, respectively. Leibniz concluded that the product uv would then change to $uv + u dv + v du$, so that $d(uv) = u dv + v du$. His argument is, roughly, that $d(uv) = (u + du)(v + dv) - uv$. Expanding the right-hand side using regular algebra and then simplifying gives $u dv + v du + du dv$. But the term $du dv$ is a second-order infinitesimal, vanishingly small compared with the first-order differentials, and is thus treated as equal to 0. Indeed, one aspect of the problems is that there appears to be an *inconsistency* in the way that infinitesimals are treated. For instance, if you want to work out the derivative of $y = x^2$, the calculation corresponding to the one just given (expanding $(x + dx)^2$, and so on) shows that $dy/dx = 2x + dx$. We then treat the dx on the right-hand side as zero, but the one on the left-hand side seems as though it ought to be an infinitesimal *nonzero* quantity, since otherwise we could not divide by it. So is it zero or not? And if not, how do we get around the apparent inconsistency?

At a slightly more technical level, the calculus required mathematicians to deal repeatedly with the “ultimate” values of ratios of the form dy/dx when the quantities in both numerator and denominator approach or actually reach 0. This phrasing uses, once again, the differential notation of Leibniz, though the same issues arose for Newton with a slightly different notational and conceptual approach. Newton generally spoke of variables as depending on time, and he sought (for example) the values approached when “evanescent increments”—vanishingly small time intervals—are considered. One long-standing set of confusions arose precisely from this idea that variable quantities were in the process of changing, whether with time or with changes in the value of another variable. This means that we talk about values of a variable approach-

ing a given value, but without a clear idea of what this “approach” actually is.

2 Eighteenth-Century Approaches and Critiques

Of course, had the calculus not turned out to be an enormously fruitful field of endeavor, no one would have bothered to criticize it. But the methods of Newton and Leibniz were widely adopted for the solution of problems that had interested earlier generations (notably tangent and area problems) and for the posing and solution of problems that these techniques suddenly made far more accessible. Problems of areas, maxima and minima, the formulation and solution of differential equations to describe the shape of hanging chains or the positions of points on vibrating strings, applications to celestial mechanics, the investigation of problems having to do with the properties of functions (thought of for the most part as analytic expressions involving variable quantities)—all these fields and more were developed over the course of the eighteenth century by such individuals as Taylor, Johann and Daniel Bernoulli, Euler, D’ALEMBERT [VI.20], Lagrange, and many others. These people employed many virtuoso arguments of suspect validity. Operations with divergent series, the use of imaginary numbers, and manipulations involving actual infinities were used effectively in the hands of the most capable of these writers. However, the methods could not always be explained to the less capable, and thus certain results were not reliably reproducible—a very odd state for mathematics from today’s standpoint. To do Euler’s calculations, one needed to be Euler. This was a situation that persisted well into the following century.

Specific controversies often highlighted issues that we now see as a result of foundational confusion. In the case of infinite series, for example, there was confusion about the domain of validity of formal expressions. Consider the series

$$1 - 1 + 1 - 1 + 1 - 1 + 1 - \dots$$

In today’s usual elementary definition (due to CAUCHY [VI.29] around 1820) we would now consider this series to be divergent because the sequence of partial sums $1, 0, 1, 0, \dots$ does not tend to a limit. But in fact there was some controversy about the actual meaning of such expressions. Euler and Nicholas Bernoulli, for example, discussed the potential distinction between the *sum* and the *value* of an infinite sum, Bernoulli arguing that something like $1 - 2 + 6 - 24 + 120 + \dots$ has no sum but

that this algebraic expression does constitute a value. Whatever may have been meant by this, Euler defended the notion that the sum of the series is the value of the finite expression that gives rise to the series. In his 1755 *Institutiones Calculi Differentialis*, he gives the example of $1 - x + x^2 - x^3 + \dots$, which comes from $1/(1+x)$, and later defended the view that this meant that $1 - 1 + 1 - 1 + \dots = \frac{1}{2}$. His view was not universally accepted. Similar controversies arose in considering how to extend the values of functions outside their usual domain, for example with the logarithms of negative numbers.

Probably the most famous eighteenth-century critique of the language and methods of eighteenth-century analysis is due to the philosopher George Berkeley (1685–1753). Berkeley’s motto, “To be is to be perceived,” expresses his idealist stance, which was coupled with a strong view that the abstraction of individual qualities, for the purposes of philosophical discussion, is impossible. The objects of philosophy should thus be things that are perceived, and perceived in their entirety. The impossibility of perceiving infinitesimally small objects, combined with their manifestly abstracted nature, led him to attack their use in his 1734 treatise *The Analyst: Or, a Discourse Addressed to an Infidel Mathematician*. Referring sarcastically in 1734 to infinitesimals as the “ghosts of departed quantities,” Berkeley argued that neglecting some quantity, no matter how small, was inappropriate in mathematical argument. He quoted Newton in this regard, to the effect that “in mathematical matters, errors are to be condemned, no matter how small.” Berkeley continued, saying that “[n]othing but the obscurity of the subject” could have induced Newton to impose this kind of reasoning on his followers. Such remarks, while they apparently did not dissuade those enamored of the methods, contributed to a sentiment that aspects of the calculus required deeper explanation. Writers such as Euler, d’Alembert, Lazare Carnot, and others attempted to address foundational criticisms by clarifying what differentials were, and gave a variety of arguments to justify the operations of the calculus.

PUP: this phrase changed, which I hope means that proofreader’s comment here has been dealt with. OK?

2.1 Euler

Euler contributed to the general development of analysis more than any other individual in the eighteenth century, and his approaches to justifying his arguments were enormously influential even after his death, owing to the success and wide use of his important textbooks.

Euler's reasoning is sometimes regarded as rather careless since he operated rather freely with the notation of the calculus, and many of his arguments are certainly deficient by later standards. This is particularly true of arguments involving infinite series and products. A typical example is provided by an early version of his proof that

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

His method is as follows. Using the known series expansion for $\sin x$ he considered the zeros of

$$\frac{\sin \sqrt{x}}{\sqrt{x}} = 1 - \frac{x}{3!} + \frac{x^2}{5!} - \frac{x^3}{7!} + \cdots.$$

These lie at π^2 , $(2\pi)^2$, $(3\pi)^2$, Applying (without argument) the factor theorem for *finite* algebraic equations he expressed this equation as

$$\left(1 - \frac{x}{\pi^2}\right) \left(1 - \frac{x}{4\pi^2}\right) \left(1 - \frac{x}{9\pi^2}\right) \cdots = 0.$$

Now, it can be seen that the coefficient of x in the infinite sum, $-\frac{1}{6}$, should equal the negative of the sum of the coefficients of x in the product. Euler apparently concluded this by imagining multiplying out the infinitely many terms and selecting the 1 from all but one of them. This gives

$$\frac{1}{\pi^2} + \frac{1}{4\pi^2} + \frac{1}{9\pi^2} + \cdots = \frac{1}{6},$$

and multiplying both sides by π^2 gives the required sum.

We now think of this approach as having several problems. The product of the infinitely many terms may or may not represent a finite value, and today we would specify conditions for when it does. Also, applying a result about (finite) polynomials to (infinite) power series is a step that requires justification. Euler himself was to provide alternative arguments for this result later in his life. But the fact that he may have known *counterexamples*—situations in which such usages would not work—was not, for him, a decisive obstacle. This view, in which one reasoned in a generic situation that might admit a few exceptions, was common at his time, and it was only in the late nineteenth century that a concerted effort was made to state the results of analysis in ways that set out precisely the conditions under which the theorems would hold.

Euler did not dwell on the interpretation of infinite sums or infinitesimals. Sometimes he was happy to regard differentials as actually equal to zero, and to

derive the meaning of a ratio of differentials from the context of the problem:

An infinitely small quantity is nothing but a vanishing quantity and therefore will be actually equal to 0. ... Hence there are not so many mysteries hidden in this concept as there are usually believed to be. These supposed mysteries have rendered the calculus of the infinitely small quite suspect to many people.

This statement, from the *Institutiones Calculi Differentialis* of 1755, was followed by a discussion of proportions in which one of the ratios is 0/0, and a justification of the fact that differentials may be neglected in calculations with ordinary numbers. This accurately describes a good deal of his practice—when he worked with differential equations, for example.

Controversial matters did arise, however, and debates about definitions were not unusual. The best-known example involves discussions connected with the so-called vibrating string problem, which involved Euler, d'Alembert, and Daniel Bernoulli. These were closely connected with the definition of FUNCTIONS [I.2 §2.2], and the question of which functions studied by analysis actually could be represented by series (in particular trigonometric series). The idea that a curve of arbitrary shape could serve as an initial position for a vibrating string extended the idea of function, and the work of FOURIER [VI.25] in the early nineteenth century made such functions analytically accessible. In this context, functions with broken graphs (a kind of *discontinuous* function) came under inspection. Later, how to deal with such functions would be a decisive issue for the foundations of analysis, as the more “natural” objects associated with algebraic operations and trigonometry gave way to the more general modern concept of function.

2.2 Responses from the Late Eighteenth Century

One significant response to Berkeley in Britain was that of Colin Maclaurin (1698–1746), whose 1742 textbook *A Treatise of Fluxions* attempted to clarify the foundations of the calculus and do away with the idea of infinitely small quantities. Maclaurin, a leading figure of the Scottish Enlightenment of the mid eighteenth century, was the most distinguished British mathematician of his time and an ardent proponent of Newton's methods. His work, unlike that of many of his British contemporaries, was read with interest on the Continent, especially his elaborations of Newtonian celestial mechanics. Maclaurin attempted to base

his reasoning on the notion of the limits of what he termed “assignable” finite quantities. Maclaurin’s work is famously obscure, though it did provide examples of calculating the limits of ratios. Perhaps his most important contribution to the clarification of the foundations of analysis was his influence on d’Alembert.

D’Alembert had read both Berkeley and Maclaurin and followed them in rejecting infinitesimals as real quantities. While exploring the idea of a differential as a limit, he also attempted to reconcile his idea with the idea that infinitesimals may be consistently regarded as being actually zero, perhaps in a nod to Euler’s view. The main exposition of d’Alembert’s views may be found in the *Encyclopédie*, in the articles on differentials (published in 1754) and on limits (1765). D’Alembert argued for the importance of geometric rather than algebraic limits. His meaning seems to have been that the quantities being investigated should not be treated merely formally, by substitution and simplification. Rather, a limit should be understood as the limit of a length (or collection of lengths), area, or other dimensioned quantity, in much the way that a circle may be seen as a limit of inscribed polygons. His aim seems primarily to have been to establish the reality of the objects described by existing algorithms, since the actual calculations he employs are carried out with differentials.

2.2.1 Lagrange

In the course of the eighteenth century, the differential and the integral calculus gradually distinguished themselves as a set of methods distinct from their applications in mechanics and physics. At the same time, the primary focus of the methods moved away from geometry, so that in work of the second half of the eighteenth century we increasingly see calculus treated as “algebraic analysis” of “analytic functions.” The term “analytic” was used in a variety of senses. For many writers, such as Euler, it merely referred to a function (that is, a relationship between variable quantities) that is given by a single expression of the type used in analysis.

Lagrange provided a foundation for the calculus that was indebted to this algebraic viewpoint. Lagrange concentrated on power-series expansions as the basic entity of analysis, and through his work the term analytic function evolved toward its more recent meaning connected with the existence of a convergent Taylor series representation. His approach reached a full expression in his *Théorie des Fonctions Analytiques* of

1797. This was a version of his lectures at the École Polytechnique, a new institution for the elite training of military engineers in revolutionary France. Lagrange assumed that a function must necessarily be expressible as an infinite series of algebraic functions, basing this argument on the existence of expansions for known functions. He first sought to show that “in general” no negative or fractional powers would appear in the expansion, and from this he obtained a power-series representation. His arguments here are surprising, and somewhat ad hoc, and I use an example given by Fraser (1987). The slightly strange notation is based on that of Lagrange. Suppose that one seeks an expansion of $f(x) = \sqrt{x+i}$ in powers of i . In general, only integer powers will be involved. Terms of the form $i^{m/n}$ do not make sense, says Lagrange, since the expression of the function $\sqrt{x+i}$ is only two-valued, while $i^{m/n}$ has n values. Hence the series

$$\sqrt{x+i} = \sqrt{x} + pi + qi^2 + \dots + ti^k + \dots$$

obtains its two values from the term \sqrt{x} , and all other powers must be integral. With fractional exponents set aside, Lagrange argued that $f(x+i) = f(x) + i^a P(x, i)$, with P finite for $i = 0$. Successive application of this result gave him the expansion

$$f(x+i) = f(x) + pi + qi^2 + ri^3 + \dots,$$

where i was a small increment. The number p depends on x , so Lagrange defined a *derived function* $f'(x) = p(x)$. The French term *dérivée* is the origin of the term derivative, and in Lagrange’s language f is the “primitive” of this derived function. Similar arguments can be made to relate the higher coefficients to the higher derivatives in the usual Taylor formula.

This approach, which seems oddly circular to modern eyes, relied on the eighteenth-century distinction between the “algebraic” infinite process of the series expansion on the one hand, and the use of differentials on the other. Lagrange did not see the original series expansion as based on the limit process. With the renewed emphasis on limits and modern definitions developed by Cauchy, this approach was soon to be regarded as untenable.

3 The First Half of the Nineteenth Century

3.1 Cauchy

Many writers contributed to discussions on rigor in analysis in the first decades of the nineteenth century. It was Cauchy who was to revive the limit approach to

greatest effect. His aim was pedagogical, and his ideas were probably worked out in the context of preparing his introductory lectures for the École Polytechnique at the beginning of the 1820s. Although the students were the best in France in scholarly ability, many found the approach too difficult. As a result, while Cauchy himself continued to use his methods, other instructors held on to older approaches using infinitesimals, which they found more intuitively accessible for the students as well as better adapted to the solution of problems in elementary mechanics. Cauchy's self-imposed exile from Paris in the 1830s further limited the impact of his approach, which was initially taken up only by a few of his students.

Nonetheless, Cauchy's definitions of limit, of continuity, and of the derivative gradually came into general use in France, and were influential elsewhere as well, especially in Italy. Moreover, his methods of using these definitions in proofs, and particularly his use of mean-value theorems in various forms, moved analysis from a collection of symbolic manipulations of quantities with special properties toward the science of argument about infinite processes using close estimation via the manipulation of inequalities.

In some respects, Cauchy's greatest contribution lay in his clear definitions. For earlier writers, the sum of an infinite series was a somewhat vague notion, sometimes interpreted by a kind of convergence argument (as with the sum of a geometric series such as $\sum_{n=0}^{\infty} 2^{-n}$) and sometimes as the value of the function from which the series was derived (as Euler, for example, often regarded it). Cauchy revised the definition to state that the sum of an infinite series was the limit of the sequence of partial sums. This provided a unified approach for series of numbers and series of functions, an important step in the move to base calculus and analysis on ideas about real numbers. This trend, eventually dominant, is often referred to as the "arithmetization of analysis." Similarly, a continuous function is one for which "an infinitely small increase of the variable produces an infinitely small increase of the function itself" (Cauchy 1821, pp. 34–35).

As we see from the example just given, Cauchy did not shy away from infinitely small quantities, nor did he analyze this notion further. The limit of a variable quantity is defined in a way that we would now regard as conversational, or heuristic:

When the values that are successively assigned to a given variable approach a fixed value indefinitely, in

such a way that it ends up differing from it as little as one wishes, this latter value is called the *limit* of all the others. Thus, for example, an irrational number is the limit of the various fractions that provide values that are closer and closer to it.

Cauchy (1821, p. 4)

These ideas were not completely rigorous by modern standards, but he was able to use them to provide a unified foundation for the basic processes of analysis.

This use of infinitely small quantities appears, for example, in his definition of a continuous function. To paraphrase his definition, suppose that a function $f(x)$ is single-valued on some finite interval of the real line, and choose any value x_0 inside the interval. If the value of x_0 is increased to $x_0 + a$, the function also changes by the amount $f(x_0 + a) - f(x_0)$. Cauchy says that the function f is continuous for this interval if, for each value of x_0 in that interval, the numerical value of the difference $f(x_0 + a) - f(x_0)$ decreases indefinitely to 0 with a . In other words, Cauchy defines continuity as a property *on an interval* rather than at a point, in essence by saying that on that interval infinitely small changes in the argument produce infinitely small changes in the function value. Cauchy appears to have considered continuity to be a property of a function on an interval.

This definition emphasizes the importance of jumps in the value of the function for the understanding of its properties, something that Cauchy had encountered early in his career when discussing THE FUNDAMENTAL THEOREM OF CALCULUS [I.3 §5.5]. In his 1814 memoir on definite integrals, Cauchy stated:

If the function $\phi(z)$ increases or decreases in a continuous manner between $z = b'$ and $z = b''$, the value of the integral $[\int_{b'}^{b''} \phi'(z) dz]$ will ordinarily be represented by $\phi(b'') - \phi(b')$. But if ... the function passes suddenly from one value to another sensibly different ... the ordinary value of the integral must be diminished.

Oeuvres (volume 1, pp. 402–3)

In his lectures, Cauchy assumed continuity when defining the definite integral. He considered first of all a division of the interval of integration into a finite number of subintervals on which the function is either increasing or decreasing. (This is not possible for all functions, but this appeared not to concern Cauchy.) He then defined the definite integral as the limit of the sum $S = (x_1 - x_0)f(x_0) + (x_2 - x_1)f(x_1) + \cdots + (X - x_{n-1})f(x_{n-1})$ as the number n becomes very

large. Cauchy gives a detailed argument for the existence of this limit, using his theorem of the mean and the fact of continuity.

Versions of the main subjects of Cauchy's lectures were published in 1821 and 1823. Every student at the École Polytechnique would have been aware of them subsequently, and many would have used them explicitly. They were joined in 1841 by a version of the course elaborated by Cauchy's associate, the Abbé Moigno. They were referred to frequently in France and the definitions employed by Cauchy became standard there. We also know that the lectures were studied by others, notably by ABEL [VI.33] and DIRICHLET [VI.36], who spent time in Paris in the 1820s, and by RIEMANN [VI.49].

Cauchy's movement away from the formal approach of Lagrange rejected the "vagueness of algebra." Although he was clearly guided by intuition (both geometric and otherwise), he was well aware that intuition could be misleading, and produced examples to show the value of adhering to precise definitions. One famous example, the function that takes the value e^{-1/x^2} when $x \neq 0$ and zero when $x = 0$, is differentiable infinitely many times, yet it does not yield a Taylor series that converges to the function at the origin. Despite this example, which he mentioned in his lectures, Cauchy was not a specialist in counterexamples, and in fact the trend toward producing counterexamples for the purpose of clarifying definitions was a later development.

Abel famously drew attention to an error in Cauchy's work: his statement that a convergent series of continuous functions has a continuous sum. For this to be true, the series must be uniformly convergent, and in 1826 Abel gave as a counterexample the series

$$\sum_{k=1}^{\infty} (-1)^{k+1} \frac{\sin kx}{k},$$

which is discontinuous at odd multiples of π . Cauchy was led to make this distinction only much later, after the phenomenon had been identified by several writers. Historians have written extensively about this apparent error; one influential account, due to Bottazzini, proposes that for various reasons Cauchy would not have found Abel's example telling, even if he had known of it at the time (Bottazzini 1990, p. LXXXV).

Before leaving the time of Cauchy, we should note the related independent activity of BOLZANO [VI.28]. Bolzano, a Bohemian priest and professor whose ideas were not widely disseminated at the time, investigated

the foundations of the calculus extensively. In 1817, for example, he gave what he termed a "purely analytic proof of the theorem that between any two values that possess opposite signs, at least one real root of the equation exists": the intermediate value theorem. Bolzano also studied infinite sets: what is now called the Bolzano–Weierstrass theorem states that in every bounded infinite set there is at least one point having the property that any disk about that point contains infinitely many points of the set. Such "limit points" were studied independently by WEIERSTRASS [VI.44]. By the 1870s, Bolzano's work became more broadly known.

3.2 Riemann, the Integral, and Counterexamples

Riemann is indelibly associated with the foundations of analysis because of the Riemann integral, which is part of every calculus course. Despite this, he was not always driven by issues involving rigor. Indeed he remains a standard example of the fruitfulness of nonrigorous intuitive invention. There are many points in Riemann's work at which issues about rigor arise naturally, and the wide interest in his innovations did much to direct the attention of researchers to making these insights precise.

Riemann's definition of the definite integral was presented in his 1854 *Habilitationschrift*—the "second thesis," which qualified him to lecture at a university for fees. He generalized Cauchy's notion to functions that are not necessarily continuous. He did this as part of an investigation of FOURIER SERIES [III.27] expansions. The extensive theory of such series was devised by Fourier in 1807 but not published until the 1820s. A Fourier series represents a function in the form

$$f(x) = a_0 + \sum_{n=1}^{\infty} (a_n \cos(nx) + b_n \sin(nx))$$

on a finite interval.

The immediate inspiration for Riemann's work was DIRICHLET [VI.36], who had corrected and developed earlier faulty work by Cauchy on the question of when and whether the Fourier series expansion of a function converges to the function from which it is derived. In 1829 Dirichlet had succeeded in proving such convergence for a function with period 2π that is integrable on an interval of that length, does not possess infinitely many maxima and minima there, and at jump discontinuities takes on the average value between the two limiting values on each side. As Riemann noted, following his professor Dirichlet, "this subject stands in the

closest connection to the principles of infinitesimal calculus, and can therefore serve to bring these to greater clarity and definiteness" (Riemann 1854, p. 238). Riemann sought to extend Dirichlet's investigations to further cases, and was thus led to investigate in detail each of the conditions given by Dirichlet. Accordingly, he generalized the definition of a definite integral as follows:

We take between a and b an increasing sequence of values x_1, x_2, \dots, x_{n-1} , and for brevity designate $x_1 - a$ by δ_1 , $x_2 - x_1$ by δ_2 , \dots , $b - x_{n-1}$ by δ_n and by ϵ a positive proper fraction. Then the value of the sum

$$S = \delta_1 f(a + \epsilon_1 \delta_1) + \delta_2 f(x_1 + \epsilon_2 \delta_2) \\ + \delta_3 f(x_2 + \epsilon_3 \delta_3) + \dots + \delta_n f(x_{n-1} + \epsilon_n \delta_n)$$

depends on the choice of the intervals δ and the quantities ϵ . If it has the property that it approaches infinitely closely a fixed limit A no matter how the δ and ϵ are chosen, as δ becomes infinitely small, then we call this value $\int_a^b f(x) dx$.

In connection with this definition of the integral, and in part to show its power, Riemann provided an example of a function that is discontinuous in any interval, yet can be integrated. The integral thus has points of nondifferentiability on each interval. Riemann's definition rendered problematic the inverse relationship between differentiation and integration, and his example brought this problem out clearly. The role of such "pathological" counterexamples in pushing the development of rigor, already apparent in Cauchy's work, intensified greatly around this time.

Riemann's definition was published only in 1867, following his death; an expository version due to Gaston Darboux appeared in French in 1873. The popularization and extension of Riemann's approach went hand in hand with the increasing appreciation of the importance of rigor associated with the Weierstrass school, discussed below. Riemann's approach focused attention on sets of points of discontinuities, and thus were seminal for CANTOR's [VI.54] investigations into point sets in the 1870s and afterwards.

The use of the *Dirichlet principle* serves as a further example of the way in which Riemann's work drew attention to problems in the foundations of analysis. In connection with his research into complex analysis, Riemann was led to investigate solutions to the so-called *Dirichlet problem*: given a function g , defined on the boundary of a closed region in the plane, does there exist a function f that satisfies the LAPLACE

PARTIAL DIFFERENTIAL EQUATION [I.3 §5.4] in the interior and takes the same values as g on the boundary? Riemann asserted that the answer was yes. To demonstrate this, he reduced the question to proving the existence of a function that minimizes a certain integral over the region, and argued on physical grounds that such a minimizing function must always exist. Even before Riemann's death his assertion was questioned by WEIERSTRASS [VI.44], who published a counterexample in 1870. This led to attempts to reformulate Riemann's results and prove them by other means, and ultimately to a rehabilitation of the Dirichlet principle through the provision of precise and broad hypotheses for its validity, which were expressed by HILBERT [VI.63] in 1900.

4 Weierstrass and His School

Weierstrass had a passion for mathematics as a student at Bonn and Münster, but his student career was very uneven. He spent the years from 1840 to 1856 as a high-school teacher, undertaking research independently but at first publishing obscurely. Papers from 1854 onward in *Journal für die reine und angewandte Mathematik* (otherwise known as *Crelle's Journal*) attracted wide attention to his talent, and he obtained a professorship in Berlin in 1856. Weierstrass began to lecture regularly on mathematical analysis, and his approach developed into a series of four courses of lectures given cyclically between the early 1860s and 1890. The lectures evolved over time and were attended by a large number of important mathematical researchers. They also indirectly influenced many others through the circulation of unpublished notes. This circle included R. Lipschitz, P. du Bois-Reymond, H. A. Schwarz, O. Hölder, Cantor, L. Koenigsberger, G. Mittag-Leffler, KOVALEVSKAYA [VI.59], and L. Fuchs, to name only some of the most important. Through their use of Weierstrassian approaches in their own research, and their espousal of his ideas in their own lectures, these approaches became widely used well before the eventual publication of a version of his lectures late in his life. The account that follows is based largely on the 1878 version of the lectures. His approach was also influential outside Germany: parts of it were absorbed in France in the lectures of HERMITE [VI.47] and JORDAN [VI.52], for example.

Weierstrass's approach builds on that of Cauchy (though the detailed relationship between the two bodies of work has never been fully examined). The two

overarching themes of Weierstrass's approach are, on the one hand, the banning of the idea of motion, or changing values of a variable, from limit processes, and, on the other, the representation of functions, notably of a complex variable. The two are intimately linked. Essential to the motion-free definition of a limit is Weierstrass's nascent investigation of what we would now call the topology of the real line or complex plane, with the idea of a limit point, and a clear distinction between local and global behavior. The central objects of study for Weierstrass are functions (of one or more real or complex variable quantities), but it should be borne in mind that set theory is not involved, so that functions are *not* to be thought of as sets of ordered pairs.

The lectures begin with a now-familiar subject: the development of rational, negative, and real numbers from the integers. For example, negative numbers are defined operationally by making the integers closed under the operation of subtraction. He attempted a unified approach to the definition of rational and irrational numbers which involved unit fractions and decimal expansions and now seems somewhat murky. While Weierstrass's definition of the real numbers appears unsatisfactory to modern eyes, the general path of *arithmetization* of analysis was established by this approach. In parallel to the development of number systems, he also developed different classes of functions, building them up from rational functions by using power-series representations. Thus, in Weierstrass's approach, a polynomial (called an integer rational function) is generalized to a "function of integer character," which means a function with a convergent power series expansion everywhere. The Weierstrass factorization theorem asserts that any such function may be written as a (possibly infinite) product of certain "prime" functions and exponential functions with polynomial exponents of a certain type.

The limit definition given by Weierstrass has thoroughly modern features:

That a variable quantity x becomes infinitely small simultaneously with another quantity y means: "After the assumption of an arbitrarily small quantity ϵ a bound δ for x may be found, such that for every value of x for which $|x| < \delta$, the corresponding value of $|y|$ will be less than ϵ ."

Weierstrass (1988, p. 57)

Weierstrass immediately used this definition to give a proof of continuity for rational functions of sev-

eral variables, using an argument that could appear in a textbook today. The former notions of variables tending to given values were replaced by quantified statements about linked inequalities. The framing of hypotheses in terms of inequalities became a guiding motif in the work of Weierstrass's school: here we mention in passing the Lipschitz and Hölder conditions in the existence theory for differential equations. The clarity that this language gave to problems involving the interchange of limits, for example, meant that previously intractable problems could now be handled in a routine way by those inculcated in the Weierstrass approach.

The fact that general functions were built from rational functions using series expansions gave the latter a key role in Weierstrass's work, and as early as 1841 he had identified the importance of uniform convergence. The distinction between uniform and pointwise convergence was made very clearly in his lectures. A series converges, as it does for Cauchy, if its sequence of partial sums converges, though now the convergence is phrased in the following terms: the series $\sum f_n(x)$ converges to s_0 at $x = x_0$ if, given an arbitrary positive ϵ , there is an integer N such that $|s_0 - (f_1(x_0) + f_2(x_0) + \cdots + f_n(x_0))| < \epsilon$ for every $n > N$. The convergence is uniform on a domain of the variable if the same N will work for that ϵ value for all x in the domain. Uniform convergence guarantees continuity of the sum, since these are series of rational, hence continuous, functions. From this point of view, then, uniform convergence is important well beyond the context of trigonometric series (important though those may be). Indeed, it is a central tool of the entire theory of functions.

Weierstrass's role as a critic of rigor in the work of others, notably Riemann, has already been noted. More than any other leading figure, he generated counterexamples to illustrate difficulties with received notions and to distinguish between different kinds of analytical behavior. One of his best-known examples was of an everywhere-continuous but nowhere-differentiable function, namely $f(x) = \sum b^n \cos(a^n x)$, which is uniformly convergent for $b < 1$ but fails to be differentiable at any x if $ab > 1 + \frac{3}{2}\pi$. Similarly he constructed functions for which the Dirichlet principle fails, examples of sets constituting "natural boundaries," that is, obstacles to continuing series expansions into larger domains, and so forth. The careful distinctions he encouraged, and the very procedure of seeking pathological rather than typical examples,

threw the spotlight on the precision of hypotheses in analysis to an unprecedented degree. From the 1880s, with the maturity of this program, analysis no longer dealt with generic cases and looked instead for absolutely precise statements in a way that has for the most part endured to the present. This was also to become a pattern and an imperative in other areas of mathematics, though sometimes the passage from reasoning from generic examples to fully expressed hypotheses and definitions took decades. (Algebraic geometry provides a famous example, one in which reasoning with generic cases lasted until the 1920s.) In this sense the form of rigorous argument and exposition espoused by Weierstrass and his school was to become a pattern for mathematics generally.

4.1 The Aftermath of Weierstrass and Riemann

Analysis became the model subdiscipline for rigor for a variety of reasons. Of course, analysis was important for the sheer volume and range of application of its results. Not everyone agreed with the precise way in which Weierstrass approached foundational questions (through series, rational functions, and so on). Indeed, Riemann's more geometric approach had attracted followers, if not exactly a school, and the insights his approach afforded were enthusiastically embraced. However, any subsequent discussion had to take place at a level of rigor comparable to that which Weierstrass had attained. While approaches to the foundations of analysis were to vary, the idea that limits should be rigorously handled in much the way that Weierstrass did was not to alter. Among the remaining central issues for rigor was the definition of the number systems.

For the real numbers, probably the most successful definition (in terms of its later use) was provided by DEDEKIND [VI.50]. Dedekind, like Weierstrass, took the integers as fundamental, and extended them to the rationals, noting that the algebraic properties satisfied by the latter are those satisfied by what we now call a FIELD [I.3 §2.2]. (This idea is also Dedekind's.) He then showed that the rational numbers satisfy a *trichotomy law*. That is, each rational number x divides the entire collection into three parts: x itself, rational numbers greater than x , and rational numbers less than x . He also showed that the rationals greater and less than a given number extend to infinity, and that any rational corresponds to a distinct point on the number line. However, he also observed that along that line there are infinitely many points that do not correspond to

any rational. Using the idea that to every point on the line there should correspond a number, he constructed the remainder of the continuum (that is, the real line) by the use of *cuts*. These are ordered pairs (A_1, A_2) of nonempty sets of rational numbers such that every element of the first set is less than every element of the second, and such that taken together they contain all the rationals. Such cuts may obviously be produced by an element x , in which case x is either the greatest element of A_1 or the least element of A_2 . But sometimes A_1 does not have a greatest element, or A_2 a least element, and in that case we can use the cut to define a new number, which is necessarily irrational. The set of all such cuts may be shown to correspond to the points of the number line, so that nothing is left out. A critical reader might feel that this is begging the question, since the idea of the number line constituting a continuum in some way might seem to be a hidden premise.

Dedekind's construction stimulated a good deal of discussion, especially in Germany, about the best way to found the real numbers. Participants included E. Heine, Cantor, and the logician FREGE [VI.56]. Heine and Cantor, for example, considered real numbers as equivalence classes of Cauchy sequences of rationals, together with a machinery that permitted them to define the basic arithmetical operations. A very similar approach was proposed by the French mathematician Charles Méray. Frege, by contrast, in his 1884 *Die Grundlagen der Arithmetik*, sought to found the integers on logic. While his attempts to construct the reals along these lines did not bear fruit, he had an important role in his insistence that the various constructions should not merely be mathematically functional but should also be demonstrably free from internal contradiction.

Despite much activity on the foundations of the real numbers, infinite sets, and other basic notions for analysis, consensus remained elusive. For example, the influential Berlin mathematician Leopold KRONECKER [VI.48] denied the existence of the reals, and held that all true mathematics was to be based on finite sets. Like Weierstrass, with whom he worked and whom he influenced, he emphasized the strong analogies between the integers and the polynomials, and sought to use this algebraic foundation to build all of mathematics. Hence for Kronecker the entire main path of research in analysis was anathema, and he opposed it ardently. These views were influential, both directly and indirectly, on a number of later writers, including BROUWER [VI.75],

the intuitionist school around him, and the algebraist and number theorist Kurt Hensel.

All efforts to found analysis were based in one way or another on an underlying notion (not always made explicit) of quantity. The foundational framework of analysis, however, was to shift over the period from 1880 to 1910 toward the theory of sets. This had its origin in the work of Cantor, a student of Weierstrass who began studying discontinuities of Fourier series in the early 1870s. Cantor became concerned about how to distinguish between different types of infinite sets. His proofs that the rational numbers and the algebraic numbers are COUNTABLE [III.11] while the reals are not led him to a hierarchy of infinite sets of different cardinality. The importance of this discovery for analysis was at first not widely recognized, though in the 1880s Mittag-Leffler and Hurwitz both made significant applications of notions about derived sets (the set of limit points of a given set) and dense or nowhere-dense sets.

Cantor gradually came to the view that set theory could function as a foundational tool for all of mathematics. As early as 1882 he wrote that the science of sets encompassed arithmetic, function theory, and geometry, combining them into a “higher unity” based on the idea of cardinality. However, this proposal was vaguely articulated and at first attracted no adherents. Nonetheless, sets began to find their way into the language of analysis, most notably through ideas of MEASURE [III.57] and measurability of a set. Indeed, one important route to the absorption of analysis by set theory was the path that sought to determine what kind of function could “measure” a set in an abstract sense. The work of LEBESGUE [VI.72] and BOREL [VI.70] around 1900 on integration and measurability tied set theory to the calculus in a very concrete and intimate way.

A further key step in the establishment of the foundations of analysis in the early twentieth century was a new emphasis on mathematical theories as axiomatic structures. This received enormous impetus from the work of Hilbert, who, beginning in the 1890s, had sought to provide a renewed axiomatization of geometry. PEANO [VI.62] in Italy headed a school with similar aims. Hilbert redefined the reals on these axiomatic grounds, and his many students and associates turned to axiomatics with enthusiasm for the clarity the approach could provide. Rather than proving the existence of specific entities such as the reals, the mathematician posits a system satisfying the fundamental properties they possess. A real number (or whatever object) is then defined by the set of axioms provided.

As Epple has pointed out, such definitions were considered to be ontologically neutral in that they did not provide methods for telling real numbers from other objects, or even state whether they existed at all (Epple 2003, p. 316). Hilbert’s student Ernst Zermelo began work on axiomatizing set theory along these lines, publishing his axioms in 1908 (see [IV.1 §3]). Problems with set theory had emerged in the form of paradoxes, the most famous due to RUSSELL [VI.71]: if S is the set of all sets that do not contain themselves, then it is not possible for S to be in S , nor can it not be in S . Zermelo’s axiomatics sought to avoid this difficulty, in part by avoiding the definition of set. By 1910, WEYL [VI.80] was to refer to mathematics as the science of “ \in ,” or set membership, rather than the science of quantity. Nonetheless, Zermelo’s axioms as a foundational strategy were contested. For one thing, a consistency proof for the axioms was lacking. Such “meaning-free” axiomatization was also contested on the grounds that it removed intuition from the picture.

Against the complex and rapidly developing background of mathematics in the early twentieth century, these debates took on many dimensions that have implications well beyond the question of what constitutes rigorous argument in analysis. For the practicing analyst, however, as well as for the teacher of basic infinitesimal calculus, these discussions are marginal to everyday mathematical life and education, and are treated as such. Set theory is pervasive in the language used to describe the basic objects. Real-valued functions of one real variable are defined as sets of ordered pairs of real numbers, for example; a set-theoretic definition of an ordered pair was given by WIENER [VI.85] in 1914, and the set-theoretic definition of functions may be dated from that time. However, research in analysis has been largely distinct from, and generally avoids, the foundational issues that may remain in connection with this vocabulary. This is not at all to say that contemporary mathematicians treat analysis in a purely formal way. The intuitive content associated with numbers and functions is very much a part of the way of thinking of most mathematicians. The axioms for the reals and for set theory form a framework to be referred to when necessary. But the essential objects of basic analysis, namely derivatives, integrals, series, and their existence or convergence behaviors, are dealt with along the lines of the early twentieth century, so that the ontological debates about the infinitesimal and infinite are no longer very lively.

A coda to this story is provided by the researches of ROBINSON [VI.95] (1918–74) into “nonstandard” analysis, published in 1961. Robinson was an expert in model theory: the study of the relationship between systems of logical axioms and the structures that may satisfy them. His differentials were obtained by adjoining to the regular real numbers a set of “differentials,” which satisfied the axioms of an ordered field (in which there is ordinary arithmetic like that of the real numbers) but in addition had elements that were smaller than $1/n$ for every positive integer n . In the eyes of some, this creation eliminated many of the unpleasant features of the usual way of dealing with the reals, and realized the ultimate goal of Leibniz to have a theory of infinitesimals which was part of the same structure as that of the reals. Despite stimulating a flurry of activity, and considerable acclaim from some quarters, Robinson’s approach has never been widely accepted as a working foundation for analysis.

Further Reading

- Bottazzini, U. 1990. Geometrical rigour and “modern analysis”: an introduction to Cauchy’s *Cours d’Analyse*. In *Cours d’Analyse de l’École Royale Polytechnique, Première Partie: Analyse Algébrique* by A.-L. Cauchy. Bologna: Editrice CLUB.
- Cauchy, A.-L. 1821. *Cours d’Analyse de l’École Royale Polytechnique, Première Partie: Analyse Algébrique*. Paris: L’Imprimerie Royale. (Reprinted, 1990, by Editrice CLUB, Bologna.)
- Epple, M. 2003. The end of the science of quantity: foundations of analysis, 1860–1910. In *A History of Analysis*, edited by H. N. Jahnke, pp. 291–323. Providence, RI: American Mathematical Society.
- Fraser, C. 1987. Joseph Louis Lagrange’s algebraic vision of the calculus. *Historia Mathematica* 14:38–53.
- Jahnke, H. N., ed. 2003. *A History of Analysis*. Providence, RI: American Mathematical Society/London Mathematical Society.
- Riemann, B. 1854. Ueber die Darstellbarkeit einer Function durch eine trigonometrische Reihe. *Königlichen Gesellschaft der Wissenschaften zu Göttingen* 13:87–131. Republished in Riemann’s collected works (1990): *Gesammelte Mathematische Werke und Wissenschaftliche Nachlass und Nachträge*, edited by R. Narasimhan, 3rd edn., pp. 259–97. Berlin: Springer.
- Weierstrass, K. 1888. *Einleitung in die Theorie der Analytischen Functionen: Vorlesung Berlin 1878*, edited by P. Ullrich. Braunschweig: Vieweg/DMV.

II.6 The Development of the Idea of Proof

Leo Corry

1 Introduction and Preliminary Considerations

In many respects the development of the idea of proof is coextensive with the development of mathematics as a whole. Looking back into the past, one might at first consider mathematics to be a body of scientific knowledge that deals with the properties of numbers, magnitudes, and figures, obtaining its justifications from proofs rather than, say, from experiments or inductive inferences. Such a characterization, however, is not without problems. For one thing, it immediately leaves out important chapters in the history of civilization that are more naturally associated with mathematics than with any other intellectual activity. For example, the Mesopotamian and Egyptian cultures developed elaborate bodies of knowledge that would most naturally be described as belonging to arithmetic or geometry, even though nothing is found in them that comes close to the idea of proof as it was later practiced in mathematics at large. To the extent that any justification is given, say, in the thousands of mathematical procedures found on clay tablets written in cuneiform script, it is inductive or based on experience. The tablets repetitively show—without additional explanation or attempts at general justifications—a given procedure to be followed whenever one is pursuing a certain type of result. Later on, in the context of Chinese, Japanese, Mayan, or Hindu cultures, one again finds important developments in fields naturally associated with mathematics. The extent to which these cultures pursued the idea of mathematical proof—a question that is debated among historians to this day—was undoubtedly not as great as it was in Greek tradition, and it certainly did not take the specific forms we typically associate with the latter. Should one nevertheless say that these are instances of mathematical knowledge, even though they are not justified on the basis of some kind of general, deductive proof? If so, then we cannot characterize mathematics as a body of knowledge that is backed up by proofs, as suggested above. However, this litmus test certainly provides a useful criterion—one that we do not want to give up too easily—for distinguishing mathematics from other intellectual endeavors.

Without totally ignoring these important questions, the present account focuses on a story that started, at some point in the past, usually taken to be before or around the fifth century B.C.E. in Greece, with the realization that there was a distinctive body of claims, mainly associated with numbers and with diagrams, whose truth could be and needed to be vindicated in a very special way—namely, by means of a general, deductive argument, or “proof.” Exactly when and how this story began is unclear. Equally unclear are the direct historical sources of such a unique idea. Since the emphasis on the use of logic and reason in constructing an argument was well-entrenched in other spheres of public life in ancient Greece—such as politics, rhetoric, and law—much earlier than the fifth century B.C.E., it is possible that it is in those domains that the origins of mathematical proof are to be found.

The early stages of this story raise some additional questions, both historical and methodological. For instance, Thales of Miletus, the first mathematician known by name (though he was also a philosopher and scientist), is reported to have *proved* several geometric theorems, such as, for instance, that the opposite angles between two intersecting straight lines are equal, or that if two vertices of a triangle are the endpoints of the diameter of a circle and the third is any other point on the circle then the triangle must be right angled. Even if we were to accept such reports at face value, several questions would immediately arise: in what sense can it be asserted that Thales “proved” these results? More specifically, what were Thales’s initial assumptions and what inference methods did he take to be valid? We know very little about this. However, we do know that, as a result of a complex historical process, a certain corpus of knowledge eventually developed that comprised known results, techniques employed, and problems (both solved and yet requiring solution). This corpus gradually also incorporated the regulatory idea of proof: that is, the idea that some kind of general argument, rather than an example (or even many examples), was the necessary justification to be sought in all cases. As part of this development, the idea of proof came to be associated with *strictly deductive* arguments, as opposed to, say, dialogic (meaning “negotiated”) or “probabilistically inferred” truth. It is an interesting and difficult historical question to establish why this was the case, and one that we will not address here.

EUCLID’s [VI.2] *Elements* was compiled some time around the year 300 B.C.E. It stands out as the most suc-

cessful and comprehensive attempt of its kind to organize the basic concepts, results, proofs, and techniques required by anyone wanting to master this increasingly complex body of knowledge. Still, it is important to stress that it was not the only such attempt within the Hellenic world. This endeavor was not just a matter of compilation, codification, and canonization, such as one can find in any other evolving field of learning at any point in time. Instead, the assertions it contained were of two different kinds, and the distinction was vitally important. On the one hand there were basic assumptions, or *axioms*, and on the other there were *theorems*, which were typically more elaborate statements, together with accounts of how they followed from the axioms—that is, proofs. The way that proof was conceived and realized in the *Elements* became the paradigm for centuries to come.

This article outlines the evolution of the idea of deductive proof as initially shaped in the framework of Euclidean-style mathematics and as subsequently practiced in the mainstream mathematical culture of ancient Greece, the Islamic world, Renaissance Europe, early modern European science, and then in the nineteenth century and at the turn of the twentieth. The main focus will be on geometry; other fields, like arithmetic and algebra, will be treated mainly in relation to it. This choice is amply justified by the subject matter itself. Indeed, much as mathematics stands out among the sciences for the unique way in which it relies on proof, so Euclidean-style geometry stood out—at least until well into the seventeenth century—among closely related disciplines such as arithmetic, algebra, and trigonometry. Individual results in these other disciplines, or indeed the domains as a whole, were often regarded as fully legitimate only when they had been provided with a geometric (or geometric-like) foundation. Important developments in nineteenth-century mathematics, mainly in connection with the rise of NON-EUCLIDEAN GEOMETRIES [II.2 §§6–10] and with problems in the FOUNDATIONS OF ANALYSIS [II.5], eventually led to a fundamental change of orientation, where arithmetic (and eventually SET THEORY [IV.1]) became the bastion of certainty and clarity from which other mathematical disciplines, geometry included, drew their legitimacy and their clarity. (See THE CRISIS IN THE FOUNDATIONS OF MATHEMATICS [II.7] for a detailed account of this development.) And yet, even before this fundamental change, Euclidean-style proof was not the only way in which mathematical proof was conceived, explored, and practiced. By focusing mainly

on geometry, the present account will necessarily leave out important developments that eventually became the mainstream of legitimate mathematical knowledge. To mention just one important example in this regard, a fundamental question that will not be pursued here is how the principle of mathematical induction originated and developed, became accepted as a legitimate inference rule of universal validity, and was finally codified as one of the basic axioms of arithmetic in the late nineteenth century. Moreover, the evolution of the notion of proof involves many other dimensions that will not be treated here, such as the development of the internal organization of mathematics into subdisciplines, as well as the changing interrelations between mathematics and its neighboring disciplines. At a different level, it is related to how mathematics itself evolved as a socially institutionalized enterprise: we shall not discuss interesting questions about how proofs are produced, made public, disseminated, criticized, and often rewritten and improved.

2 Greek Mathematics

Euclid's *Elements* is the paradigmatic work of Greek mathematics, partly for what it has to say about the basic concepts, tools, results, and problems of synthetic geometry and arithmetic, but also for how it regards the role of a mathematical proof and the form that such a proof takes. All proofs appearing in the *Elements* have six parts and are accompanied by a diagram. I illustrate this with the example of proposition I.37. Euclid's text is quoted here in the classical translation of Sir Thomas Heath, and the meaning of some terms differs from current usage. Thus, two triangles are said to be "in the same parallels" if they have the same height and both their bases are contained in a single line, and any two figures are said to be "equal" if their areas are equal. For the sake of explanation, names of the parts of the proof have been added: these do not appear in the original. The proof is illustrated in figure 1.

Protasis (enunciation). Triangles which are on the same base and in the same parallels are equal to one another.

Ekthesis (setting out). Let ABC, DBC be triangles on the same base BC and in the same parallels AD, BC.

Diorismos (definition of goal). I say that the triangle ABC is equal to the triangle DBC.

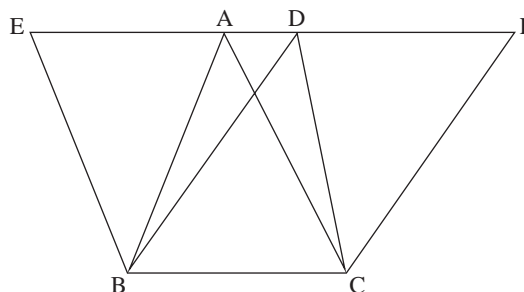


Figure 1 Proposition I.37 of Euclid's *Elements*.

Kataskeue (construction). Let AD be produced in both directions to E, F; through B let BE be drawn parallel to CA, and through C let CF be drawn parallel to BD.

Apodeixis (proof). Then each of the figures EBCA, DBFC is a parallelogram; and they are equal, for they are on the same base BC and in the same parallels BC, EF. Moreover the triangle ABC is half of the parallelogram EBCA, for the diameter AB bisects it; and the triangle DBC is half of the parallelogram DBCF, for the diameter DC bisects it. Therefore the triangle ABC is equal to the triangle DBC.

Sumperasma (conclusion). Therefore triangles which are on the same base and in the same parallels are equal to one another.

This is an example of a proposition that states a property of geometric figures. The *Elements* also includes propositions that express a task to be carried out. An example is proposition I.1: "On a given finite straight line to construct an equilateral triangle." The same six parts of the proof and the diagram invariably appear in propositions of this kind as well. This formal structure is also followed in all propositions appearing in the three *arithmetic* books of the *Elements* and, most importantly, all of them are always accompanied by a diagram. Thus, for instance, consider proposition IX.35, which in its original version reads as follows:

If as many numbers as we please be in continued proportion, and there be subtracted from the second and the last numbers equal to the first, then, as the excess of the second is to the first, so will the excess of the last be to all those before it.

This cumbersome formulation may prove incomprehensible on first reading. In more modern terms, an equivalent to this theorem would state that, given a geometric progression a_1, a_2, \dots, a_{n+1} , we have

$$(a_{n+1} - a_1) : (a_1 + a_2 + \dots + a_n) = (a_2 - a_1) : a_1.$$

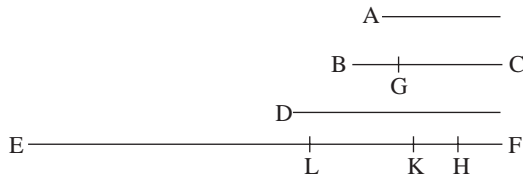


Figure 2 Proposition IX.35 of Euclid's *Elements*.

This translation, however, fails to convey the spirit of the original, in which no formal symbolic manipulation is, or can be, made. More importantly, a modern algebraic proof fails to convey the ubiquity of diagrams in Greek mathematical proofs, even where they are not needed for a truly geometric construction. Indeed, the accompanying diagram for proposition IX.35 is shown as figure 2 and the first few lines of the proof are as follows:

Let there be as many numbers as we please in continued proportion A, BC, D, EF , beginning from A as least and let there be subtracted from BC and EF the numbers BG, FH , each equal to A ; I say that, as GC is to A , so is EH to A, BC, D . For let FK be made equal to BC and FL equal to D ...

This proposition and its proof provide good examples of the capabilities, as well as the limitations, of ancient Greek practices of notation, and especially of how they managed without a truly symbolic language. In particular, they demonstrate that proofs were never conceived by the Greeks, even ideally, as purely logical constructs, but rather as specific kinds of arguments that one applied to a diagram. The diagram was not just a visual aid to the argumentation. Rather, through the *ekthesis* part of the proof, it embodied the idea referred to by the general character and formulation of the proposition.

Together with the centrality of diagrams, the six-part structure is also typical of most of Greek mathematics. The constructions and diagrams that typically appeared in Greek mathematical proofs were not of an arbitrary kind, but what we identify today as straightedge-and-compass constructions. The reasoning in the *apodeixis* part could be either a direct deduction or an argument by contradiction, but the result was always known in advance and the proof was a means to justify it. In addition, Greek geometric thinking, and in particular Euclid-style geometric proofs, strictly adhered to a principle of homogeneity. That is, magnitudes were only compared with, added to, or subtracted

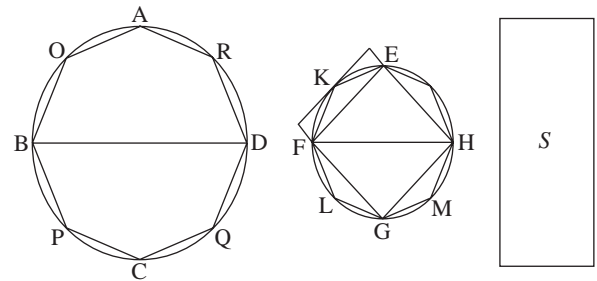


Figure 3 Proposition XII.2 of Euclid's *Elements*.

from magnitudes of like kind—numbers, lengths, areas, or volumes. (See NUMBERS [II.1 §2] for more about this.)

Of particular interest are those Greek proofs concerned with lengths of curves, as well as with areas or volumes enclosed by curvilinear shapes. Greek mathematicians lacked a flexible notation capable of expressing the gradual approximation of curves by polygons and an eventual passage to the infinite. Instead, they devised a special kind of proof that involved what can retrospectively be seen as an implicit passage to the limit, but which did so in the framework of a purely geometric proof and thus unmistakably followed the six-part proof-scheme described above. This implicit passage to the infinite was based on the application of a continuity principle, later associated with ARCHIMEDES [VI.3]. In Euclid's formulation, for instance, the principle states that, given two unequal magnitudes of the same kind, A, B (be they two lengths, two areas, or two volumes), with A greater than B , and if we subtract from A a magnitude which is greater than $A/2$, and from the remainder we subtract a magnitude that is greater than its half, and if this process is iterated a sufficient number of times, then we will eventually remain with a magnitude that is smaller than B . Euclid used this principle to prove, for instance, that the ratio of the areas of two circles equals the ratio of the squares of their diameters (XII.2). The method used, later known as the *exhaustion method*, was based on a *double contradiction* that became standard for many centuries to come. This double contradiction is illustrated in figure 3, the accompanying diagram to the proposition.

If the ratio of the square on BD to the square on FH is not the same as the ratio of circle $ABCD$ to circle $EFGH$, then it must be the same as the ratio of circle $ABCD$ to an area S either larger or smaller than circle $EFGH$. The curvilinear figures are approximated by polygons, since the continuity principle allows the difference between the inscribed polygon and the circle

to be as close as desired (e.g., closer than the difference between S and EFGH). The “double contradiction” is reached if one assumes that S is either smaller or larger than EFGH.

Forms of proof and constructions other than those mentioned so far are occasionally found in Greek mathematical texts. These include diagrams based on what is assumed to be the synchronized motion of two lines (e.g., the trisectrix, or Archimedes’ spiral), mechanical devices of many sorts, or reasoning based on idealized mechanical considerations. However, the Euclidean type of proof described above remained a model to be followed wherever possible. There is a famous Archimedes palimpsest that provides evidence of how less canonical methods, drawing on mechanical considerations (albeit of a highly idealized kind), were used to deduce results about areas and volumes. However, even this bears testimony to the primacy of the ideal model: there is a letter from Archimedes to Eratosthenes in which he displays the ingenuity of his mechanical methods but at the same time is at pains to stress their heuristic character.

3 Islamic and Renaissance Mathematics

Just as Euclid is considered to be representative of a mainstream tradition in Greek mathematics, AL-KHWĀRIZMĪ [VI.5] is regarded as a typical representative of Islamic mathematics. There are two main traits of his work that are relevant to the present account and that became increasingly central to the development of mathematics, starting with his works in the late eighth century and continuing until the works of CARDANO [VI.7] in sixteenth-century Italy. These traits are a pervasive “algebraization” of mathematical thinking, and a continued reliance on Euclidean-style geometric proof as the main way of legitimizing the validity of mathematical knowledge in general and of algebraic reasoning in mathematics in particular.

The prime example of this combination is found in al-Khwārizmī’s seminal text *al-Kitāb al-mukhtaṣar fī ḥisāb al-jabr wa’l-muqābala* (“The compendious book on calculation by completion and balancing”), where he discusses the solutions of problems in which the unknown length appears in combination with numbers and squares (the side of which is an unknown). Since he only envisages the possibility of positive “coefficients” and positive rational solutions, al-Khwārizmī needs to consider six different situations each of which requires

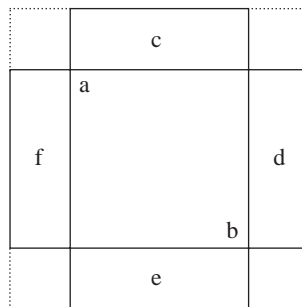


Figure 4 Al-Khwārizmī’s geometric justification of the formula for a quadratic equation.

a different recipe for finding the unknown: the full-grown idea of a general quadratic equation and an algorithm to solve it in all cases does not appear in Islamic mathematical texts. For instance, the problem “squares and roots equal to numbers” (e.g., $x^2 + 10x = 39$, in modern notation) and the problem “roots and numbers equal to squares” (e.g., $3x + 4 = x^2$) are considered to be completely different ones, as are their solutions, and accordingly al-Khwārizmī treats them separately. In all cases, however, al-Khwārizmī *proves* the validity of the method described by translating it into geometric terms and then relying on Euclid-like geometric theorems built around a specific diagram. It is noteworthy, however, that the problems refer to specific numerical quantities associated with the magnitudes involved, and these measured magnitudes refer to the accompanying diagrams as well. In this way, al-Khwārizmī interestingly departs from the Euclidean style of proof. Still, the Greek principle of homogeneity is essentially preserved, as the three quantities usually involved in the problem are all of the same kind, namely, areas.

Consider, for instance, the equation $x^2 + 10x = 39$, which corresponds to the following problem of al-Khwārizmī.

What is the square which combined with ten of its roots will give a sum total of 39?

The recipe prescribes the following steps.

Take one-half of the roots [5] and multiply them by itself [25]. Add this amount to 39 and obtain 64. Take the square root of this, which is eight, subtract from it half the roots, leaving three. The number three therefore represents one root of this square, which itself, of course, is nine.

The *justification* is provided by figure 4.

Here ab represents the said square, which for us is x^2 , and the rectangles c, d, e, f represent an area of $\frac{10}{4}x$ each, so that all of them together equal $10x$, as in the problem. Thus, the small squares in the corners represent an area of 6.25 each, and we can “complete” the large square, being equal to 64, and whose side is therefore 8, thus yielding the solution 3 for the unknown.

Abu Kamil Shuja, just one generation after al-Khwārizmī, added force to this approach when he solved additional problems while specifically relying on theorems taken from the *Elements*, including the accompanying diagrams, in order to justify his method of solution. The primacy of the Euclidean-type proof, which was already accepted in geometry and arithmetic, thus also became associated with the algebraic methods that eventually turned into the main topic of interest in Renaissance mathematics. Cardano’s 1545 *Ars Magna*, the foremost example of this new trend, presented a complete treatment of the equations of third and fourth degree. Although the algebraic line of reasoning that he adopted and developed became increasingly abstract and formal, Cardano continued to justify his arguments and methods of solution by reference to Euclid-like geometric arguments based on diagrams.

4 Seventeenth-Century Mathematics

The next significant change in the conception of proof appears in the seventeenth century. The most influential development of mathematics in this period was the creation of the infinitesimal calculus simultaneously by NEWTON [VI.14] and LEIBNIZ [VI.15]. This momentous development was the culmination of a process that spanned most of the century, involving the introduction and gradual improvement of important techniques for determining areas and volumes, gradients of tangents, and maxima and minima. These developments included the elaboration of traditional points of view that went back to the Greek classics, as well as the introduction of completely new ideas such as the “indivisibles,” whose status as a legitimate tool for mathematical proof was hotly debated. At the same time, the algebraic techniques and approaches that Renaissance mathematicians continued to expand upon, following on from their Islamic predecessors, now gained additional impetus and were gradually incorporated—starting with the work of FERMAT [VI.12] and DESCARTES [VI.11]—into the arsenal of tools available for

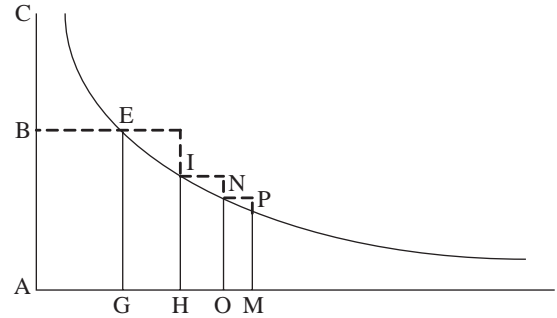


Figure 5 Diagram for Fermat’s proof of the area under a hyperbola.

proving geometric results. Underlying these various trends were different conceptions and practices of mathematical proof, which are briefly described and illustrated now.

Examples of how the classical Greek conception of geometric proof was essentially followed but at the same time fruitfully modified and expanded are found in the work of Fermat, as can be seen in his calculation of the area enclosed by a generalized hyperbola (in modern notation $(y/a)^m = (x/b)^n$ ($m, n \neq 1$)) and its asymptotes.

The quadratic hyperbola (i.e., a figure represented by $y = 1/x^2$), for instance, is defined here in terms of a purely geometric relationship on any two of its points, namely, that the ratio between the squares built on the abscissas equals the inverse ratio between the lengths of the ordinates. In its original version it is expressed as follows: $AG^2 : AH^2 :: IH : EG$ (see figure 5). It should be noticed that this is not an equation in the present sense of the word, on which the standard symbolic manipulations can be directly performed. Rather, this is a four-term proportion to which the rules of Greek classical mathematics apply. Also, the proof was entirely geometric and indeed it essentially followed the Euclidean style. Thus, if the segments AG, AH, AO , etc., are chosen in continued proportion, then one can prove that the rectangles EH, IO, NM , etc., are also in continued proportion, and indeed that $EH : IO :: IO : NM :: \dots :: AH : AG$.

Fermat made use of proposition IX.35 of the *Elements* (mentioned above), which comprises an expression for the sum of any number of quantities in a geometric progression, namely (in more modern notation):

$$(a_{n+1} - a_1) : (a_1 + a_2 + \dots + a_n) = (a_2 - a_1) : a_1.$$

But at this point his proof takes an interesting turn. He introduces the somewhat obscure concept of “*adequare*,” which he found in the works of Diophantus, and which allows a kind of “approximate equality.” Specifically, this idea allows him to bypass the cumbersome procedure of double contradiction typically used in Greek geometry as an implicit passage to the infinite. A figure bounded by GE, by the horizontal asymptote, and by the hyperbola will equal the infinite sum of rectangles obtained when the rectangle EH “will vanish and will be reduced to nothing.” Further, proposition IX.35 implies that this sum equals the area of the rectangle BG. Significantly, Fermat still chose to rely on the authority of the ancients, hinting at the method of double contradiction when he declared that this result “would be easy to confirm by a more lengthy proof carried out in the manner of Archimedes.”

Attempts to expand the accepted canon of geometric proof eventually led to the more progressive approaches associated with the idea of indivisibles (described below), as practiced by Cavalieri, Roberval, and Torricelli. This is well-illustrated by Torricelli’s 1643 calculation of the volume of the infinite body created by (expressed in modern terms) rotating the hyperbola $xy = k^2$ around the y -axis, with values of x between 0 and a .

The essential idea of indivisibles is that areas are considered to be sums, or collections, of infinitely many line segments, and volumes are considered to be sums, or collections, of infinitely many areas. In this example, Torricelli calculated the volume of revolution by considering it to be a sum of the curved surfaces of an infinite collection of cylinders successively inscribed within each other and having radii ranging from 0 to a . The area of the curved surface of the inscribed cylinder with radius x is $2\pi x(k^2/x)$ and is thus equal, for any x , to the area of the circle AS, where S is the point (k, k) on the hyperbola in figure 6(b).

However, from the figure it can be seen that in building the entire rotational body there is a cylindrical surface associated with each possible length between 0 and a , and therefore that the total volume of the infinite body can be considered as being composed of all the cylindrical surfaces, which in turn equals the infinite sum of circles, each of which is associated with a radius between 0 and a (see figure 6(c)), and which is equal to the volume of a cylinder with radius AS and height a (see figure 6(d)).

The rules of Euclid-like geometric proof were completely contravened in proofs of this kind and this

made them unacceptable in the eyes of many. On the other hand, their fruitfulness was highly appealing, especially in cases like this one in which an infinite body was shown to have a finite volume, a result which Torricelli himself found extremely surprising. Both supporters and detractors alike, however, realized that techniques of this kind might lead to contradictions and inaccurate results. By the eighteenth century, with the accelerated development of the infinitesimal calculus and its associated techniques and concepts, techniques based on indivisibles had essentially disappeared.

The limits set by the classical paradigm of Euclidean geometric proof were then transgressed in a different direction by the all-embracing algebraization of geometry at the hands of Descartes. The fundamental step undertaken by Descartes was to introduce unit lengths as a key element in the diagrams used in geometric proofs. The radical innovation implied by this step, allowing the hitherto nonexistent possibility of defining operations with line segments, was explicitly stressed by Descartes in *La Géométrie* in 1637:

Just as arithmetic consists of only four or five operations, namely addition, subtraction, multiplication, division, and the extraction of roots, which may be considered a kind of division, so in geometry, to find required lines it is merely necessary to add or subtract other lines; or else, taking one line, which I shall call the unit in order to relate it as closely as possible to numbers, and which can in general be chosen arbitrarily, and having given two other lines, to find a fourth line which shall be to one of the given lines as the other is to the unit (which is the same as multiplication); or again, to find a fourth line which is to one of the given lines as the unit is to the other (which is equivalent to division); or, finally, to find one, two, or several mean proportionals between the unit and some other line (which is the same as extracting the square root, cube root, etc., of the given line).

Thus, for instance, given two segments BD, BE, the division of their lengths is represented by BC in figure 7, in which AB represents the unit length.

Although the proof was Euclid-like in appearance (because of the diagram and the use of the theory of similar triangles), the introduction of the unit length and its use for defining the operations with segments set it radically apart and opened completely new horizons for geometric proofs. Not only had measurements of length been absent from Euclidean-style proofs thus far, but also, as a consequence of the very existence of these operations, the essential dimensionality traditionally associated with geometric theorems lost its

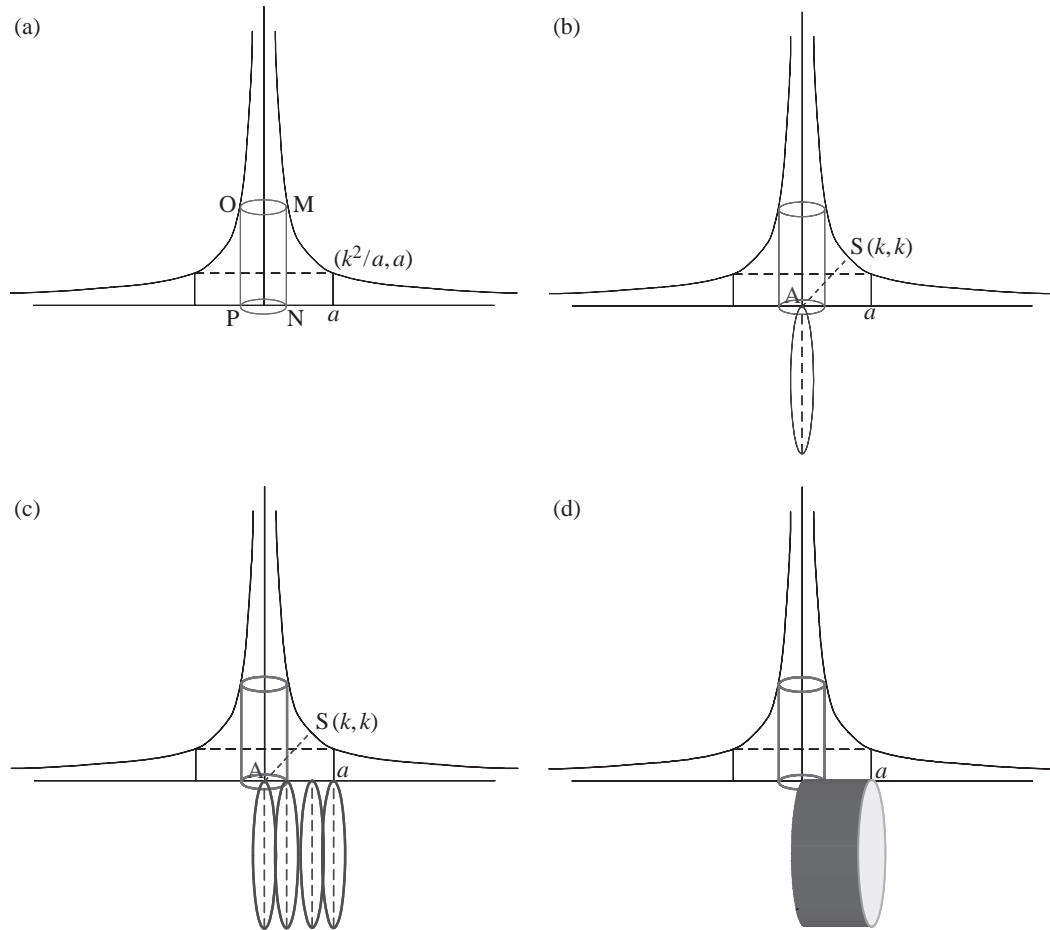


Figure 6 Torricelli's proof of the volume of an infinite body.

significance. Descartes used expressions such as $a - b$, a/b , a^2 , b^3 , and their roots, but he stressed that they should all be understood as “only simple lines, which, however, I name squares, cubes, etc., so that I make use of the terms employed in algebra.” With the removal of dimensionality, the requirement of homogeneity also became unnecessary. Unlike his predecessors, who handled magnitudes only when they had a direct geometric significance, Descartes could not see any problem in forming an expression such as $a^2b^2 - b$ and then extracting its cube root. In order to do so, he said “we must consider the quantity a^2b^2 divided once by the unit, and the quantity b multiplied twice by the unit.” Sentences of this kind would be simply incomprehensible to Greek geometers, as well as to their Islamic and Renaissance followers.

This algebraization of geometry, and particularly the newly created possibility of proving geometric facts via algebraic procedures, was strongly related to the recent consolidation of the idea of an algebraic equation, seen as an autonomous mathematical entity, for which formal rules of manipulation were well-known and could be systematically applied. This idea reached full maturity in the hands of VIÈTE [VI.9] only around 1591. But not all mathematicians in the seventeenth century saw the important developments associated with algebraic thinking either as a direction to be naturally adopted or as a clear sign of progress in the latter discipline. A prominent opponent of any attempt to deviate from the classical Euclidean-style approach in geometry was none other than NEWTON [VI.14], who, in the *Arithmetica Universalis* (1707), was emphatic in expressing his views:

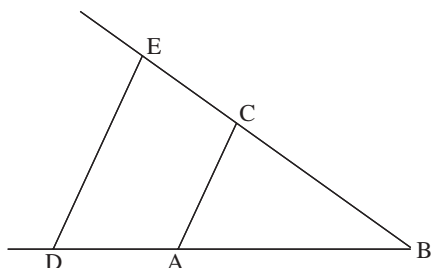


Figure 7 Descartes's geometric calculation of the division of two given segments.

Equations are expressions of arithmetic computation and properly have no place in geometry, except in so far as truly geometrical quantities (lines, surfaces, solids and proportions) are thereby shown equal, some to others. Multiplications, divisions, and computations of that kind have recently been introduced into geometry, unadvisedly and against the first principle of this science.... Therefore these two sciences ought not to be confounded, and recent generations by confounding them have lost that simplicity in which all geometrical elegance consists.

Newton's *Principia* bears witness to the fact that statements like this one were far from mere lip service, as Newton consistently preferred Euclidean-style proofs, considering them to be the correct language for presenting his new physics and for bestowing it with the highest degree of certainty. He used his own calculus only where strictly necessary, and barred algebra from his treatise entirely.

5 Geometry and Proof in Eighteenth-Century Mathematics

Mathematical analysis became the primary focus of mathematicians in the eighteenth century. Questions relating to the foundations of analysis arose immediately after the calculus began to be developed and were not settled until the late nineteenth century. To a considerable extent these questions were about the nature of legitimate mathematical proof, and debates about them played an important role in undermining the long-undisputed status of geometry as the basis for mathematical certainty and bestowing this status on arithmetic instead. The first important stage in this process was EULER's [VI.19] reformulation of the calculus. Once separated from its purely geometric roots, the calculus came to be centered on the algebraically oriented concept of function. This trend for favoring algebra over

geometry was given further impetus by Euler's successors. D'ALEMBERT [VI.20], for instance, associated mathematical certainty above all with algebra—because of its higher degree of generality and abstraction—and only subsequently with geometry and mechanics. This was a clear departure from the typical views of Newton and of his contemporaries. The trend reached a peak and was transformed into a well-conceived program in the hands of LAGRANGE [VI.22], who in the preface to his 1788 *Mécanique Analytique* famously expressed a radical view about how one could achieve certainty in the mathematical sciences while distancing oneself from geometry. He wrote as follows:

One will not find figures in this work. The methods that I expound require neither constructions, nor geometrical or mechanical arguments, but only algebraic operations, subject to a regular and uniform course.

The details of these developments are beyond the scope of this article. What is important to stress, however, is that in spite of their very considerable impact, the basic conceptions of proof in the more mainstream realm of *geometry* did not change very much during the eighteenth century. An illuminating perspective on these conceptions is offered by the views of contemporary philosophers, especially Immanuel Kant.

Kant had a very profound knowledge of contemporary science, and particularly of mathematics. A *philosophical* discussion of his views on mathematical knowledge and proof need not concern us here. However, given his acquaintance with contemporary conceptions, they do provide an insightful *historical* perspective on proof as it was understood at the time. Of particular interest is the contrast he draws between a philosophical argument, on the one hand, and a geometric proof, on the other. Whereas the former deals with general concepts, the latter deals with concrete, yet nonempirical, concepts, by reference to “visualizable intuitions” (*Anschauung*). This difference is epitomized in the following, famous passage from his *Critique of Pure Reason*.

Suppose a philosopher be given the concept of a triangle and he is left to find out, in his own way, what relation the sum of its angles bears to a right angle. He has nothing but the concept of a figure enclosed by three straight lines, and possessing three angles. However long he meditates on this concept, he will never produce anything new. He can analyze and clarify the concept of a straight line or of an angle or of the number three, but he can never arrive at any properties not

already contained in these concepts. Now let the geometrician take up these questions. He at once begins by constructing a triangle. Since he knows that the sum of two right angles is exactly equal to the sum of all the adjacent angles which can be constructed from a single point on a straight line, he prolongs one side of his triangle and obtains two adjacent angles, which together are equal to two right angles. He then divides the external angle by drawing a line parallel to the opposite side of the triangle, and observes that he has thus obtained an external adjacent angle which is equal to an internal angle—and so on. In this fashion, through a chain of inferences guided throughout by intuition, he arrives at a fully evident and universally valid solution of the problem.

In a nutshell, then, for Kant the nature of mathematical proof that sets it apart from other kinds of deductive argumentation (like philosophy) lies in the centrality of the diagrams and the role that they play. As in the *Elements*, this diagram is not just a heuristic guide for what is no more than abstract reasoning, but rather an “intuition,” a singular embodiment of the mathematical idea that is clearly located not only in space, but rather in space and time. In fact,

I cannot represent to myself a line, however small, without drawing it in thought, that is gradually generating all its parts from a point. Only in this way can the intuition be obtained.

This role played by diagrams as “visualizable intuitions” is what provides, for Kant, the explanation of why geometry is not just an empirical science, but also not just a huge tautology devoid of any synthetic content. According to him, geometric proof is constrained by logic but it is much more than just a purely logical analysis of the terms involved. This view was at the heart of a novel philosophical analysis whose starting point was the then-entrenched conception of what a mathematical proof is.

6 Nineteenth-Century Mathematics and the Formal Conception of Proof

The nineteenth century was full of important developments in geometry and other parts of mathematics, not just of the methods but also of the aims of the various subdisciplines. Logic, as a field of knowledge, also underwent significant changes and a gradual mathematization that entirely transformed its scope and methods. Consequently, by the end of the century the conception of proof and its role in mathematics had also been deeply transformed.

In Göttingen in 1854 RIEMANN [VI.49] gave his seminal talk “On the hypotheses which lie at the foundations of geometry.” At around the same time, the works of BOLYAI [VI.34] and LOBACHEVSKII [VI.31] on non-Euclidean geometry, as well as the related ideas of GAUSS [VI.26], all dating from the 1830s, began to be more generally known. The existence of coherent, alternative geometries brought about a pressing need for the most basic, longstanding beliefs about the essence of geometric knowledge, including the role of proof and mathematical rigor, to be revised. Of even greater significance in this regard was the renewed interest in PROJECTIVE GEOMETRY [I.3 §6.7], which became a very active field of research with its own open research questions and foundational issues after the publication of Jean Poncelet’s 1822 treatise. The addition of projective geometry to the many other possible geometric perspectives prompted a variety of attempts at unification and classification, the most significant of which were those based on group-theoretic ideas. Particularly notable were those of KLEIN [VI.57] and LIE [VI.53] in the 1870s. In 1882, Moritz Pasch published an influential treatise on projective geometry devoted to a systematic exploration of its axiomatic foundations and the interrelationships among its fundamental theorems. Pasch’s book also attempted to close the many logical gaps that had been found in Euclidean geometry over the years. More systematically than any of his fellow nineteenth-century mathematicians, Pasch emphasized that all geometric results should be obtained from axioms by strict logical deduction, without relying on analytical means, and above all without appeal to diagrams or to properties of the figures involved. Thus, although in some ways he was consciously reverting to the canons of Euclid-like proof (which by then were somewhat loosened), his attitude toward diagrams was fundamentally different. Aware of the potential limitations of visualizing diagrams (and perhaps their misleading influence) he put a much greater emphasis on the pure logical structure of the proof than his predecessors had. Nevertheless, he was not led to an outright formalist view of geometry and geometric proof. Rather, he consistently adopted an empirical approach to the origins and meaning of geometry and fell short of claiming that diagrams were for heuristic use only:

The basic propositions [of geometry] cannot be understood without corresponding drawings; they express what has been observed from certain, very simple facts. The theorems are not founded on observations, but

rather, they are proved. Every inference performed during a deduction must find confirmation in a drawing, yet it is not justified by a drawing but from a certain preceding statement (or a definition).

Pasch's work definitely contributed to diagrams losing their central status in geometric proofs in favor of purely deductive relations, but it did not directly lead to a thorough revision of the status of the axioms of geometry, or to a change in the conception that geometry deals essentially with the study of our spatial, visualizable intuition (in the sense of *Anschauung*). The all-important nineteenth-century developments in geometry produced significant changes in the conception of proof only under the combined influence of additional factors.

Mathematical analysis continued to be a primary field of research, and the study of its foundations became increasingly identified with arithmetic, rather than geometric, rigor. This shift was provoked by the works of mathematicians like CAUCHY [VI.29], WEIERSTRASS [VI.44], CANTOR [VI.54], and DEDEKIND [VI.50], which aimed at eliminating intuitive arguments and concepts in favor of ever more elementary statements and definitions. (In fact, it was not until the work of Dedekind on the foundations of arithmetic, in the last third of the century, that the rigorous formulation pursued in these works was given any kind of axiomatic underpinning.) The idea of investigating the axiomatic basis of mathematical theories, whether geometry, algebra, or arithmetic, and of exploring alternative possible systems of postulates was indeed pursued during the nineteenth century by mathematicians such as George Peacock, Charles Babbage, John Herschel, and, in a different geographical and mathematical context, Hermann Grassmann. But such investigations were the exception rather than the rule, and they had only a fairly limited role in shaping a new conception of proof in analysis and geometry.

One major turning point, where the above trends combined to produce a new kind of approach to proof, is to be found in the works of Giuseppe PEANO [VI.62] and his Italian followers. Peano's mainstream activities were as a competent analyst, but he was also interested in artificial languages, and particularly in developing an artificial language that would allow a completely formal treatment of mathematical proofs. In 1889 his successful application of such a conceptual language to arithmetic yielded his famous POSTULATES FOR THE NATURAL NUMBERS [III.69]. Pasch's systems of axioms for

projective geometry posed a challenge to Peano's artificial language, and he set out to investigate the relationship between the logical and the geometric terms involved in the deductive structure of geometry. In this context he introduced the idea of an independent set of axioms, and applied this concept to his own system of axioms for projective geometry, which were a slight modification of Pasch's. This view did not lead Peano to a formalistic conception of proof, and he still conceived geometry in terms very similar to his predecessors:

Anyone is allowed to take a hypothesis and develop its logical consequences. However, if one wants to give this work the name of geometry it is necessary that such hypotheses or postulates express the result of simple and elementary observations of physical figures.

Under the influence of Peano, Mario Pieri developed a symbolism with which to handle abstract-formal theories. Unlike Peano and Pasch, Pieri consistently promoted the idea of geometry as a purely logical system, where theorems are deduced from hypothetical premises and where the basic terms are completely detached from any empirical or intuitive significance.

A new chapter in the history of geometry and of proof was opened at the end of the nineteenth century with the publication of HILBERT's [VI.63] *Grundlagen der Geometrie*, a work that synthesized and brought to completion the various trends of geometric research described above. Hilbert was able to achieve a comprehensive analysis of the logical interrelations among the fundamental results of projective geometry, such as the theorems of Desargues and Pappus, while paying particular attention to the role of continuity considerations within their proofs. His analysis was based on the introduction of a generalized analytic geometry, in which the coordinates may be taken from a variety of different NUMBER FIELDS [III.65], rather than from the real numbers alone. This approach created a purely synthetic arithmetization of any given type of geometry, and thus helped to clarify the logical structure of Euclidean geometry as a deductive system. It also clarified the relationship between Euclidean geometry and the various other kinds of known geometries—non-Euclidean, projective, or non-Archimedean. This focus on logic implied, among other things, that diagrams should be relegated to a merely heuristic role. In fact, although diagrams still appear in many proofs in the *Grundlagen*, the entire purpose of the logical analysis is to avoid being misled by diagrams. Proofs, and partic-

ularly geometric proofs, have thus become purely logical arguments, rather than arguments about diagrams. And at the same time, the essence and the role of the axioms from which the derivations in question start also underwent a dramatic change.

Following Pasch's lead, Hilbert introduced a new system of axioms for geometry that attempted to close the logical gaps inherent in earlier systems. These axioms were of five kinds—axioms of incidence, of order, of congruence, of parallels, and of continuity—each of which expressed a particular way in which spatial intuition manifests itself in our understanding. They were formulated for three fundamental kinds of object: points, lines, and planes. These remained undefined, and the system of axioms was meant to provide an implicit definition of them. In other words, rather than defining points or lines at the outset and then postulating axioms that are assumed to be valid for them, a point and a line were not directly defined, except as entities that satisfy the axioms postulated by the system. Further, Hilbert demanded that the axioms in a system of this kind should be mutually independent, and introduced a method for checking that this demand is fulfilled; in order to do so, he constructed models of geometries that fail to satisfy a given axiom of the system but satisfy all the others. Hilbert also required that the system be consistent, and that the consistency of geometry could be made to depend, in his system, on that of arithmetic. He initially assumed that proving the consistency of arithmetic would not present a major obstacle and it was a long time before he realized that this was not the case. Two additional requirements that Hilbert initially introduced for axiomatic systems were simplicity and completeness. Simplicity meant, in essence, that an axiom should not contain more than "a single idea." The demand that every axiom in a system be "simple," however, was never clearly defined or systematically pursued in subsequent works of Hilbert or any of his successors. The last requirement, completeness, meant for Hilbert in 1900 that any adequate axiomatization of a mathematical domain should allow for a derivation of *all* the known theorems of the discipline in question. Hilbert claimed that his axioms would indeed yield all the known results of Euclidean geometry, but of course this was not a property that he could formally prove. In fact, since this property of "completeness" cannot be formally checked for any given axiomatic system, it did not become one of the standard requirements of an axiomatic system. It is important to note that the concept of completeness used by

Hilbert in 1900 is completely different from the currently accepted, model-theoretical one that appeared much later. The latter amounts to the requirement that in a given axiomatic system every true statement, be it known or unknown, should be provable.

The use of undefined concepts and the concomitant conception of axioms as implicit definitions gave enormous impetus to the view of geometry as a purely logical system, such as Pieri had devised it, and eventually transformed the very idea of truth and proof in mathematics. Hilbert claimed on various occasions—echoing an idea of Dedekind—that, in his system, "points, lines, and planes" could be substituted by "chairs, tables, and beer mugs," without thereby affecting in any sense the logical structure of the theory. Moreover, in the light of discussions about set-theoretical paradoxes, Hilbert strongly emphasized the view that the logical consistency of a concept implicitly defined by axioms was the essence of mathematical existence. Under the influence of these views, of the new methodological tools introduced by Hilbert, and of the successful overview of the foundations of geometry thus achieved, many mathematicians went on to promote new views of mathematics and new mathematical activities that in many senses went beyond the views embodied in Hilbert's approach. On the one hand, a trend that thrived in the United States at the beginning of the twentieth century, led by Eliakim H. Moore, turned the study of systems of postulates into a mathematical field in its own right, independent of direct interest in the field of research defined by the systems in question. For instance, these mathematicians defined the minimal set of independent postulates for groups, fields, projective geometry, etc., without then proceeding to investigate any of these individual disciplines. On the other hand, prominent mathematicians started to adopt and develop increasingly formalistic views of proof and of mathematical truth, and began applying them in a growing number of mathematical fields. The work of the radically modernist mathematician Felix HAUSDORFF [VI.68] provides important examples of this trend, as he was among the first to consistently associate Hilbert's achievement with a new, formalistic view of geometry. In 1904, for instance, he wrote:

In all philosophical debates since Kant, mathematics, or at least geometry, has always been treated as heteronomous, as dependent on some external instance of what we could call, for want of a better term, intuition, be it pure or empirical, subjective or scientifically amended, innate or acquired. The most important and

fundamental task of modern mathematics has been to set itself free from this dependency, to fight its way through from heteronomy to autonomy.

Hilbert himself would pursue such a point of view around 1918, when he engaged in the debates about the consistency of arithmetic and formulated his “finitist” program. This program did indeed adopt a strongly formalistic view, but it did so with the restricted aim of solving this particular problem. It is therefore important to stress that Hilbert’s conceptions of geometry were, and remained, essentially empiricist and that he never regarded his axiomatic analysis of geometry as part of an overall formalistic conception of mathematics. He considered the axiomatic approach as a tool for the conceptual clarification of existing, well-elaborated theories, of which geometry provided only the most prominent example.

The implication of Hilbert’s axiomatic approach for the concept of proof and of truth in mathematics provoked strong reactions from some mathematicians, and prominently so from FREGE [VI.56]. Frege’s views are closely connected with the changing status of logic at the turn of the twentieth century and its gradual process of mathematization and formalization. This process was an outcome of the successive efforts through the nineteenth century of BOOLE [VI.43], DE MORGAN [VI.38], Grassmann, Charles S. Peirce, and Ernst Schröder at formulating an algebra of logic. The most significant step toward a new, formal conception of logic, however, came with the increased understanding of the role of the logical QUANTIFIERS [I.2 §3.2] (universal, \forall , and existential, \exists) in the process of formulating a modern mathematical proof. This understanding emerged in an informal, but increasingly clear, fashion as part of the process of the rigorization of analysis and the distancing from visual intuition, especially at the hands of Cauchy, BOLZANO [VI.28], and Weierstrass. It was formally defined and systematically codified for the first time by Frege in his 1879 *Begriffsschrift*. Frege’s system, as well as similar ones proposed later by Peano and by RUSSELL [VI.71], brought to the fore a clear distinction between propositional connectives and quantifiers, as well as between logical symbols and algebraic or arithmetic ones.

Frege formulated the idea of a *formal system*, in which one defines in advance all the allowable symbols, all the rules that produce well-formed formulas, all axioms (i.e., certain preselected, well-formed formulas), and all the rules of inference. In such systems

any deduction can be checked *syntactically*—in other words, by purely symbolic means. On the basis of such systems Frege aimed to produce theories with no logical gaps in their proofs. This would apply not only to analysis and to its arithmetic foundation—the mathematical fields that provided the original motivation for his work—but also to the new systems of geometry that were evolving at the time. On the other hand, in Frege’s view the axioms of mathematical theories—even if they appear in the formal system merely as well-formed formulas—embody truths about the world. This is precisely the source of his criticism of Hilbert. It is the truth of the axioms, asserted Frege, that certifies their consistency, rather than the other way around, as Hilbert suggested.

We thus see how foundational research in two separate fields—geometry and analysis—was inspired by different methodologies and philosophical outlooks, but converged at the turn of the twentieth century to create an entirely new conception of mathematical proof. In this conception a mathematical proof is seen as a purely logical construct validated in purely syntactic terms, independently of any visualization through diagrams. This conception has dominated mathematics ever since.

Epilogue: Proof in the Twentieth Century

The new notion of proof that stabilized at the beginning of the twentieth century provided an idealized model—broadly accepted to this day—of what should constitute a valid mathematical argument. To be sure, actual proofs devised and published by mathematicians since that time are seldom presented as fully formalized texts. They typically present a clearly articulated argument in a language that is precise enough to convince the reader that it could—in principle, and perhaps with straightforward (if sustained) effort—be turned into one. Throughout the decades, however, some limitations of this dominant idea have gradually emerged and alternative conceptions of what should count as a valid mathematical argument have become increasingly accepted as part of current mathematical practice.

The attempt to pursue this idea systematically to its full extent led, early on and very unexpectedly, to a serious difficulty with the notion of a proof as a completely formalized and purely syntactic deductive argument. In the early 1920s, Hilbert and his collaborators developed a fully fledged mathematical theory whose subject matter was “proof,” considered as an object of

study in itself. This theory, which presupposed the formal conception of proof, arose as part of an ambitious program for providing a direct, *finitistic* consistency proof of arithmetic represented as a formalized system. Hilbert asserted that, just as the physicist examines the physical apparatus with which he carries out his experiments and the philosopher engages in a critique of reason, so the mathematician should be able to analyze mathematical proofs and do so strictly by mathematical means. About a decade after the program was launched, GÖDEL [VI.92] came up with his astonishing INCOMPLETENESS THEOREM [V.18], which famously showed that “mathematical truth” and “provability” were not one and the same thing. Indeed, in any consistent, sufficiently rich axiomatic system (including the systems typically used by mathematicians) there are true mathematical statements that cannot be proved. Gödel’s work implied that Hilbert’s finitistic program was too optimistic, but at the same time it also made clear the deep mathematical insights that could be obtained from Hilbert’s proof theory.

A closely related development was the emergence of proofs that certain important mathematical statements were undecidable. Interestingly, these seemingly negative results have given rise to new ideas about the legitimate grounds for establishing the truth of such statements. For instance, in 1963 Paul Cohen established that the CONTINUUM HYPOTHESIS [IV.1 §5] can be neither proved nor disproved in the usual systems of axioms for set theory. Most mathematicians simply accept this idea and regard the problem as solved (even if not in the way that was originally expected), but some contemporary set theorists, notably Hugh Woodin, maintain that there are good reasons to believe that the hypothesis is *false*. The strategy they follow in order to justify this assertion is fundamentally different from the formal notion of proof: they devise new axioms, demonstrate that these axioms have very desirable properties, argue that they should therefore be accepted, and then show that they imply the negation of the continuum hypothesis. (See SET THEORY [IV.1 §10] for further discussion.)

A second important challenge came from the ever-increasing length of significant proofs appearing in various mathematical domains. A prominent example was the CLASSIFICATION THEOREM FOR FINITE SIMPLE GROUPS [V.8], whose proof was worked out in many separate parts by a large numbers of mathematicians. The resulting arguments, if put together, would reach about ten thousand pages, and errors have been found since

the announcement in the early 1980s that the proof was complete. It has always been relatively straightforward to fix the errors and the theorem is indeed accepted and used by group theorists. Nevertheless, the notion of a proof that is too long for a single human being to check is a challenge to our conception of when a proof should be accepted as such. The more recent, very conspicuous cases of FERMAT’S LAST THEOREM [V.12] and THE POINCARÉ CONJECTURE [V.28] were hard to survey for different reasons: not only were they long (though nowhere near as long as the classification of finite simple groups), but they were also very difficult. In both cases there was a significant interval between the first announcement of the proofs and their complete acceptance by the mathematical community because checking them required enormous efforts by the very few people qualified to do so. There is no controversy about either of these two breakthroughs, but they do raise an interesting sociological problem: if somebody claims to have proved a theorem and nobody else is prepared to check it carefully (perhaps because, unlike the two theorems just mentioned, this one is not important enough for another mathematician to be prepared to spend the time that it would take), then what is the status of the theorem?

Proofs based on probabilistic considerations have also appeared in various mathematical domains, including number theory, group theory, and combinatorics. It is sometimes possible to prove mathematical statements (see, for example, the discussion of random primality testing in COMPUTATIONAL NUMBER THEORY [IV.5 §2]), not with complete certainty, but in such a way that the probability of error is tiny—at most one in a trillion, say. In such cases, we may not have a formal proof, but the chances that we are mistaken in considering the given statement to be true are probably lower than, say, than the chance that there is a significant mistake in one of the lengthy proofs mentioned above.

Another challenge has come from the introduction of computer-assisted methods of proof. For instance, in 1976 Kenneth Appel and Wolfgang Haken settled a famous old problem by proving the FOUR-COLOR THEOREM [V.14]. Their proof involved the checking of a huge number of different map configurations, which they did with the help of a computer. Initially, this raised debates about the legitimacy of their proof but it quickly became accepted and there are now several proofs of this kind. Some mathematicians even believe that computer-assisted and, more importantly, *computer-generated* proofs are the future of the entire

discipline. Under this (currently minority) view, our present views about what counts as an acceptable mathematical proof will soon become obsolete.

A last point to stress is that many branches of mathematics now contain conjectures that seem to be both fundamentally important and out of reach for the foreseeable future. Mathematicians persuaded of the truth of such conjectures increasingly undertake the systematic study of their consequences, assuming that an acceptable proof will one day appear (or at least that the conjecture is true). Such conditional results are published in leading mathematical journals and doctoral degrees are routinely awarded for them.

All of these trends raise interesting questions about existing conceptions of legitimate mathematical proofs, the status of truth in mathematics, and the relationship between “pure” and “applied” fields. The formal notion of a proof as a string of symbols that obeys certain syntactical rules continues to provide an ideal model for the principles that underlie what most mathematicians see as the essence of their discipline. It allows far-reaching mathematical analysis of the power of certain axiomatic systems, but at the same time it falls short of explaining the changing ways in which mathematicians decide what kinds of arguments they are willing to accept as legitimate in their actual professional practice.

I thank José Ferreirós and Reviel Netz for useful comments on previous versions of this text.

Further Reading

- Bos, H. 2001. *Redefining Geometrical Exactness. Descartes' Transformation of the Early Modern Concept of Construction*. New York: Springer.
- Ferreirós, J. 2000. *Labyrinth of Thought. A History of Set Theory and Its Role in Modern Mathematics*. Boston, MA: Birkhäuser.
- Grattan-Guinness, I. 2000. *The Search for Mathematical Roots, 1870–1940: Logics, Set Theories and the Foundations of Mathematics from Cantor through Russell to Gödel*. Princeton, NJ: Princeton University Press.
- Netz, R. 1999. *The Shaping of Deduction in Greek Mathematics: A Study in Cognitive History*. Cambridge: Cambridge University Press.
- Rashed, R. 1994. *The Development of Arabic Mathematics: Between Arithmetic and Algebra*, translated by A. F. W. Armstrong. Dordrecht: Kluwer.

II.7 The Crisis in the Foundations of Mathematics

José Ferreirós

The foundational crisis is a celebrated affair among mathematicians and it has also reached a large non-mathematical audience. A well-trained mathematician is supposed to know something about the three viewpoints called “logicism,” “formalism,” and “intuitionism” (to be explained below), and about what GÖDEL’S INCOMPLETENESS RESULTS [V.18] tell us about the status of mathematical knowledge. Professional mathematicians tend to be rather opinionated about such topics, either dismissing the foundational discussion as irrelevant—and thus siding with the winning party—or defending, either as a matter of principle or as an intriguing option, some form of revisionist approach to mathematics. But the real outlines of the historical debate are not well-known and the subtler philosophical issues at stake are often ignored. Here we shall mainly discuss the former, in the hope that this will help bring the main conceptual issues into sharper focus.

The foundational crisis is usually understood as a relatively localized event in the 1920s, a heated debate between the partisans of “classical” (meaning late-nineteenth-century) mathematics, led by HILBERT [VI.63], and their critics, led by BROUWER [VI.75], who advocated strong revision of the received doctrines. There is, however, a second, and in my opinion very important, sense in which the “crisis” was a long and global process, indistinguishable from the rise of modern mathematics and the philosophical and methodological issues it created. This is the standpoint from which the present account has been written.

Within this longer process one can still pick out some noteworthy intervals. Around 1870 there were many discussions about the acceptability of non-Euclidean geometries, and also about the proper foundations of complex analysis and even the real numbers. Early in the twentieth century there were debates about set theory, about the concept of the continuum, and about the role of logic and the axiomatic method versus the role of intuition. By about 1925 there was a crisis in the proper sense, during which the main opinions in these debates were developed and turned into detailed mathematical research projects. And in the 1930s GÖDEL [VI.92] proved his incompleteness results,

which could not be assimilated without some cherished beliefs being abandoned. Let us analyze some of these events and issues in greater detail.

1 Early Foundational Questions

There is evidence that in 1899 Hilbert endorsed the viewpoint that came to be known as *logicism*. Logicism was the thesis that the basic concepts of mathematics are definable by means of logical notions, and that the key principles of mathematics are deducible from logical principles alone.

Over time this thesis has become unclear, based as it seems to be on a fuzzy and immature conception of the scope of logical theory. But historically speaking logicism was a neat intellectual reaction to the rise of modern mathematics, and particularly to the set-theoretic approach and methods. Since the majority opinion was that set theory is just a part of (refined) logic,¹ this thesis was thought to be supported by the fact that the theories of natural and real numbers can be derived from set theory, and also by the increasingly important role of set-theoretic methods in algebra and in real and complex analysis.

Hilbert was following DEDEKIND [VI.50] in the way he understood mathematics. For us, the essence of Hilbert's and Dedekind's early logicism is their self-conscious endorsement of certain modern methods, however daring they seemed at the time. These methods had emerged gradually during the nineteenth century, and were particularly associated with Göttingen mathematics (GAUSS [VI.26] and DIRICHLET [VI.36]); they experienced a crucial turning point with RIEMANN's [VI.49] novel ideas, and were developed further by Dedekind, CANTOR [VI.54], Hilbert, and other, lesser figures. Meanwhile, the influential Berlin school of mathematics had opposed this new trend, KRONECKER [VI.48] head-on and WEIERSTRASS [VI.44] more subtly. (The name of Weierstrass is synonymous with the introduction of rigor in real analysis, but in fact, as will be indicated below, he did not favor the more modern methods elaborated in his time.) Mathematicians in Paris and elsewhere also harbored doubts about these new and radical ideas.

The most characteristic traits of the modern approach were:

- (i) acceptance of the notion of an "arbitrary" function proposed by Dirichlet;
- (ii) a wholehearted acceptance of infinite sets and the higher infinite;
- (iii) a preference "to put thoughts in the place of calculations" (Dirichlet), and to concentrate on "structures" characterized axiomatically; and
- (iv) a reliance on "purely existential" methods of proof.

An early and influential example of these traits was Dedekind's approach (1871) to ALGEBRAIC NUMBER THEORY [IV.3]—his set-theoretic definition of NUMBER FIELDS [III.65] and IDEALS [III.83 §2], and the methods by which he proved results such as the fundamental theorem of unique decomposition. In a remarkable departure from the number-theoretic tradition, Dedekind studied the factorization properties of algebraic integers in terms of ideals, which are certain infinite sets of algebraic integers. Using this new abstract concept, together with a suitable definition of the product of two ideals, Dedekind was able to prove in full generality that, within any ring of algebraic integers, ideals possess a unique decomposition into prime ideals.

The influential algebraist Kronecker complained that Dedekind's proofs do not enable us to calculate, in a particular case, the relevant divisors or ideals: that is, the proof was *purely existential*. Kronecker's view was that this abstract way of working, made possible by the set-theoretic methods and by a concentration on the algebraic properties of the structures involved, was too remote from an algorithmic treatment—that is, from so-called *constructive* methods. But for Dedekind this complaint was misguided: it merely showed that he had succeeded in elaborating the principle "to put thoughts in the place of calculations," a principle that was also emphasized in Riemann's theory of complex functions. Obviously, concrete problems would require the development of more delicate computational techniques, and Dedekind contributed to this in several papers. But he also insisted on the importance of a general, conceptual theory.

The ideas and methods of Riemann and Dedekind became better known through publications of the period 1867–72. These were found particularly shocking because of their very explicit defense of the view that mathematical theories *ought not* to be based upon formulas and calculations—they should always be based on clearly formulated *general concepts*, with

1. One should mention that key figures like Riemann and Cantor disagreed (see Ferreirós 1999). The "majority" included Dedekind, PEANO [VI.62], Hilbert, RUSSELL [VI.71], and others.

analytical expressions or calculating devices relegated to the further development of the theory.

To explain the contrast, let us consider the particularly clear case of the opposition between Riemann's and Weierstrass's approaches to function theory. Weierstrass opted systematically for explicit representations of analytic (or HOLOMORPHIC [I.3 §5.6]) functions by means of power series of the form $\sum_{n=0}^{\infty} a_n(z-a)^n$, connected with each other by ANALYTIC CONTINUATION [I.3 §5.6]. Riemann chose a very different and more abstract approach, defining a function to be analytic if it satisfies the CAUCHY-RIEMANN DIFFERENTIABILITY CONDITIONS [I.3 §5.6].² This neat conceptual definition appeared objectionable to Weierstrass, as the class of differentiable functions had never been carefully characterized (in terms of series representations, for example). Exercising his famous critical abilities, Weierstrass offered examples of continuous functions that were nowhere differentiable.

It is worth mentioning that, in preferring infinite series as the key means for research in analysis and function theory, Weierstrass remained closer to the old eighteenth-century idea of a function as an analytical expression. On the other hand, Riemann and Dedekind were always in favor of Dirichlet's abstract idea of a function f as an "arbitrary" way of associating with each x some $y = f(x)$. (Previously it had been required that y should be expressed in terms of x by means of an explicit formula.) In his letters, Weierstrass criticized this conception of Dirichlet's as too general and vague to constitute the starting point for any interesting mathematical development. He seems to have missed the point that it was in fact just the right framework in which to define and analyze general concepts such as CONTINUITY [I.3 §5.2] and INTEGRATION [I.3 §5.5]. This framework came to be called the *conceptual approach* in nineteenth-century mathematics.

Similar methodological debates emerged in other areas too. In a letter of 1870, Kronecker went as far as saying that the Bolzano-Weierstrass theorem was an "obvious sophism," promising that he would offer counterexamples. The Bolzano-Weierstrass theorem, which states that an infinite bounded set of real numbers has an accumulation point, was a cornerstone

of classical analysis, and was emphasized as such by Weierstrass in his famous Berlin lectures. The problem for Kronecker was that this theorem rests entirely on the completeness axiom for the real numbers (which, in one version, states that every sequence of nonempty nested closed intervals in \mathbb{R} has a nonempty intersection). The real numbers cannot be constructed in an elementary way from the rational numbers: one has to make heavy use of infinite sets (such as the set of all possible "Dedekind cuts," which are subsets $C \subset \mathbb{Q}$ such that $p \in C$ whenever p and q are rational numbers such that $p < q$ and $q \in C$). To put it another way: Kronecker was drawing attention to the problem that, very often, the accumulation point in the Bolzano-Weierstrass theorem cannot be constructed by elementary operations from the rational numbers. The classical idea of the set of real numbers, or "the continuum," already contained the seeds of the *nonconstructive* ingredient in modern mathematics.

Later on, in around 1890, Hilbert's work on invariant theory led to a debate about his purely existential proof of another basic result, the *basis theorem*, which states (in modern terminology) that every ideal in a polynomial ring is finitely generated. Paul Gordan, famous as the "king" of invariants for his heavily algorithmic work on the topic, remarked humorously that this was "theology," not mathematics! (He apparently meant that, because the proof was purely existential, rather than constructive, it was comparable with philosophical proofs of the existence of God.)

This early foundational debate led to a gradual clarification of the opposing viewpoints. Cantor's proofs in set theory also became quintessential examples of the modern methodology of existential proof. He offered an explicit defense of the higher infinite and modern methods in a paper of 1883, which was peppered with hidden attacks on Kronecker's views. Kronecker in turn criticized Dedekind's methods publicly in 1882, spoke privately against Cantor, and in 1887 published an attempt to spell out his foundational views. Dedekind replied with a detailed set-theoretic (and "thus," for him, logicistic) theory of the natural numbers in 1888.

The early round of criticism ended with an apparent victory for the modern camp, which enrolled new and powerful allies such as Hurwitz, MINKOWSKI [VI.64], Hilbert, Volterra, Peano, and HADAMARD [VI.65], and which was defended by influential figures such as KLEIN [VI.57]. Although Riemannian function theory was still in need of further refinement, recent developments in real analysis, number theory, and other fields were

2. Riemann determined particular functions by a series of *independent* traits such as the associated RIEMANN SURFACE [III.81] and the behavior at singular points. These traits determined the function via a certain variational principle (the "Dirichlet principle"), which was also criticized by Weierstrass, who gave a counterexample to it. Hilbert and Kneser would later reformulate and justify the principle.

showing the power and promise of the modern methods. During the 1890s, the modern viewpoint in general, and logicism in particular, enjoyed great expansion. Hilbert developed the new methodology into the axiomatic method, which he used to good effect in his treatment of geometry (1899 and subsequent editions) and of the real number system.

Then, dramatically, came the so-called logical paradoxes, discovered by Cantor, Russell, Zermelo, and others, which will be discussed below. These were of two kinds. On the one hand, there were arguments showing that assumptions that certain sets exist lead to contradictions. These were later called the *set-theoretic* paradoxes. On the other, there were arguments, later known as the *semantic* paradoxes, which showed up difficulties with the notions of truth and definability. These paradoxes completely destroyed the attractive view of recent developments in mathematics that had been proposed by logicism. Indeed, the heyday of logicism came *before* the paradoxes, that is, before 1900; it subsequently enjoyed a revival with Russell and his “theory of types,” but by 1920 logicism was of interest more to philosophers than to mathematicians. However, the divide between advocates of the modern methods and constructivist critics of these methods was there to stay.

2 Around 1900

Hilbert opened his famous list of mathematical problems at the Paris International Congress of Mathematics of 1900 with Cantor’s CONTINUUM PROBLEM [IV.1 §5], a key question in set theory, and with the problem of whether every set can be well-ordered. His second problem amounted to establishing the consistency of the notion of the set \mathbb{R} of real numbers. It was not by chance that he began with these problems: rather, it was a way of making a clear statement about how mathematics should be in the twentieth century. Those two problems, and THE AXIOM OF CHOICE [III.1] employed by Hilbert’s young colleague Zermelo to show that \mathbb{R} (the *continuum*) can be well-ordered, are quintessential examples of the traits (i)–(iv) that were listed above. It is little wonder that less daring minds objected and revived Kronecker’s doubts, as can be seen in many publications of 1905–6. This brings us to the next stage of the debate.

2.1 Paradoxes and Consistency

In a remarkable turn of events, the champions of modern mathematics stumbled upon arguments that cast new doubts on its cogency. In around 1896, Cantor discovered that the seemingly harmless concepts of the set of all ordinals and the set of all cardinals led to contradictions. In the former case the contradiction is usually called the *Burali-Forti paradox*; the latter is the *Cantor paradox*. The assumption that all transfinite ordinals form a set leads, by Cantor’s previous results, to the result that there is an ordinal that is less than itself—and similarly for cardinals. Upon learning of these paradoxes, Dedekind began to doubt whether human thought is completely rational. Even worse, in 1901–2 Zermelo and Russell discovered a very elementary contradiction, now known as *Russell’s paradox* or sometimes as the *Zermelo–Russell paradox*, which will be discussed in a moment. The untenability of the previous understanding of set theory as logic became clear, and there began a new period of instability. But it should be said that only logicians were seriously upset by these arguments: they were presented with contradictions in their theories.

Let us explain the importance of the Zermelo–Russell paradox. From Riemann to Hilbert, many authors accepted the principle that, given any well-defined logical or mathematical property, there exists a set of *all* objects satisfying that property. In symbols: given a well-defined property p , there exists another object, the set $\{x : p(x)\}$. For example, corresponding to the property of “being a real number” (which is expressed formally by Hilbert’s axioms) there is the set of all real numbers; corresponding to the property of “being an ordinal” there is the set of all ordinals; and so on. This is called the *comprehension principle*, and it constitutes the basis for the logicistic understanding of set theory, often called naive set theory, although its naivete is only clear with hindsight. The principle was thought of as a basic logical law, so that all of set theory was merely a part of elementary logic.

The Zermelo–Russell paradox shows that the comprehension principle is contradictory, and it does so by formulating a property that seems to be as basic and purely logical as possible. Let $p(x)$ be the property $x \notin x$ (bearing in mind that negation and membership were assumed to be purely logical concepts). The comprehension principle yields the existence of the set $R = \{x : x \notin x\}$, but this leads quickly to a contradiction: if $R \in R$, then $R \notin R$ (by the definition

of R), and similarly, if $R \notin R$, then $R \in R$. Hilbert (like his older colleague FREGE [VI.56]) was led to abandon logicism, and even wondered whether Kronecker might have been right all along. Eventually he concluded that set theory had shown the need to refine logical theory. It was also necessary to establish set theory axiomatically, as a basic *mathematical* theory based on mathematical (not logical) axioms, and Zermelo undertook this task.

Hilbert famously advocated that to claim that a set of mathematical objects exists is tantamount to proving that the corresponding axiom system is consistent—that is, free of contradictions. The documentary evidence suggests that Hilbert came to this celebrated principle in reaction to Cantor's paradoxes. His reasoning may have been that, instead of jumping directly from well-defined concepts to their corresponding sets, one had first to prove that the concepts are logically consistent. For example, before one could accept the set of all real numbers, one should prove the consistency of Hilbert's axiom system for them. Hilbert's principle was a way of removing any metaphysical content from the notion of mathematical existence. This view, that mathematical objects had a sort of "ideal existence" in the realm of thought rather than an independent metaphysical existence, had been anticipated by Dedekind and Cantor.

The "logical" paradoxes included not only the ones that go by the names of Burali-Forti, Cantor, and Russell, but also many semantic paradoxes formulated by Russell, Richard, König, Grelling, etc. (Richard's paradox will be discussed below.) Much confusion emerged from the abundance of different paradoxes, but one thing is clear: they played an important role in promoting the development of modern logic and convincing mathematicians of the need for strictly formal presentation of their theories. Only when a theory has been stated within a precise formal language can one disregard the semantic paradoxes, and even formulate the distinction between these and the set-theoretic ones.

2.2 Predicativity

When the books of Frege and Russell made the paradoxes of set theory widely known to the mathematical community in 1903, POINCARÉ [VI.61] used them to put forward criticisms of both logicism and formalism.

His analysis of the paradoxes led him to coin an important new notion, *predicativity*, and maintain that impredicative definitions should be avoided in mathematics. Informally, a definition is impredicative when

it introduces an element by reference to a totality that already contains that element. A typical example is the following: Dedekind defines the set \mathbb{N} of natural numbers as the intersection of all sets that contain 1 and are closed under an injective function σ such that $1 \notin \sigma(\mathbb{N})$. (The function σ is called the *successor function*.) His idea was to characterize \mathbb{N} as minimal, but in his procedure the set \mathbb{N} is first introduced by appeal to a totality of sets that should already include \mathbb{N} itself. This kind of procedure appeared unacceptable to Poincaré (and also to Russell), especially when the relevant object can be specified *only* by reference to the more embracing totality. Poincaré found examples of impredicative procedures in each of the paradoxes he studied.

Take, for instance, Richard's paradox, which is one of the linguistic or semantic paradoxes (where, as we said, the notions of truth and definability are prominent). One begins with the idea of *definable* real numbers. Because definitions must be expressed in a certain language by finite expressions, there are only countably many definable numbers. Indeed, we can explicitly count the definable real numbers by listing them in alphabetical order of their definitions. (This is known as the *lexicographic order*.) Richard's idea was to apply to this list a diagonal process, of the kind used by Cantor to prove that \mathbb{R} is not COUNTABLE [III.11]. Let the definable numbers be a_1, a_2, a_3, \dots . Define a new number r in a systematic way, making sure that the n th decimal digit of r is different from the n th decimal digit of a_n . (For example, let the n th digit of r be 2 unless the n th digit of a_n is 2, in which case let the n th digit of r be 4.) Then r cannot belong to the set of definable numbers. But in the course of this construction, the number r has just been defined in finitely many words! Poincaré would ban impredicative definitions and would therefore prevent the introduction of the number r , since it was defined with reference to the totality of all definable numbers.³

In this kind of approach to the foundations of mathematics, all mathematical objects (beyond the natural numbers) must be introduced by explicit definitions. If a definition refers to a presumed totality of which the object being defined is itself a member, we are involved in a circle: the object itself is then a constituent of its own definition. In this view, "definitions"

3. The modern solution is to establish mathematical definitions within a well-determined formal theory, whose language and expressions are fixed to begin with. Richard's paradox takes advantage of an ambiguity as to what the available means of definition are.

must be predicative: one refers only to totalities that have already been established before the object one is defining. Important authors such as Russell and WEYL [VI.80] accepted this point of view and developed it.

Zermelo was not convinced, arguing that impredicative definitions were often used unproblematically, not only in set theory (as in Dedekind's definition of \mathbb{N} , for example), but also in classical analysis. As a particular example, he cited CAUCHY's [VI.29] proof of THE FUNDAMENTAL THEOREM OF ALGEBRA [V.15],⁴ but a simpler example of impredicative definition is the least upper bound in real analysis. The real numbers are not introduced separately, by explicit predicative definitions of each one of them; rather, they are introduced as a completed whole, and the particular way in which the least upper bound of an infinite bounded set of reals is singled out becomes impredicative. But Zermelo insisted that these definitions are innocuous, because the object being defined is not "created" by the definition; it is merely singled out (see his paper of 1908 in van Heijenoort (1967, pp. 183–98)).

Poincaré's idea of abolishing impredicative definitions became important for Russell, who incorporated it as the "vicious circle principle" in his influential *theory of types*. Type theory is a system of higher-order logic, with quantification over properties or sets, over relations, over sets of sets, and so on. Roughly speaking, it is based on the idea that the elements of any set should always be objects of a certain homogeneous type. For instance, we can have sets of "individuals," such as $\{a, b\}$, or sets of *sets* of individuals, such as $\{\{a\}, \{a, b\}\}$, but never a "mixed" set like $\{a, \{a, b\}\}$. Russell's version of type theory became rather complicated because of the so-called ramification he adopted in order to avoid impredicativity. This system, together with axioms of infinity, choice, and "reducibility" (a surprisingly ad hoc means to "collapse" the ramification), sufficed for the development of set theory and the number systems. Thus it became the logical basis for the renowned *Principia Mathematica* by Whitehead and Russell (1910–13), in which they carefully developed a foundation for mathematics.

Type theory remained the main logical system until about 1930, but under the form of *simple* type theory

(that is, without ramification), which, as Chwistek, Ramsey, and others realized, suffices for a foundation in the style of *Principia*. Ramsey proposed arguments that were aimed at eliminating worries about impredicativity, and he tried to justify the other existence axioms of *Principia*—the axiom of infinity and the axiom of choice—as logical principles. But his arguments were inconclusive. Russell's attempt to rescue logicism from the paradoxes remained unconvincing, except to some philosophers (especially members of the Vienna Circle).

Poincaré's suggestions also became a key principle for the interesting foundational approach proposed by Weyl in his book *Das Kontinuum* (1918). The main idea was to accept the theory of the natural numbers as they were conventionally developed using classical logic, but to work predicatively from there on. Thus, unlike Brouwer, Weyl accepted the principle of the excluded middle. (This, and Brouwer's views, will be discussed in the next section.) However, the full system of the real numbers was not available to him: in his system the set \mathbb{R} was not complete and the Bolzano–Weierstrass theorem failed, which meant that he had to devise sophisticated replacements for the usual derivations of results in analysis.

The idea of *predicative foundations* for mathematics, in the style of Weyl, has been carefully developed in recent decades with noteworthy results (see Feferman 1998). Predicative systems lie between those that countenance all of the modern methodology and the more stringent constructivistic systems. This is one of several foundational approaches that do not fit into the conventional but by now outdated triad of logicism, formalism, and intuitionism.

2.3 Choices

As important as the paradoxes were, their impact on the foundational debate has often been overstated. One frequently finds accounts that take the paradoxes as the real starting point of the debate, in strong contrast with our discussion in section 1. But even if we restrict our attention to the first decade of the twentieth century, there was another controversy of equal, if not greater, importance: the arguments that surrounded the axiom of choice and Zermelo's proof of the well-ordering theorem.

Recall from section 2.1 that the association between sets and their defining properties was at the time deeply ingrained in the minds of mathematicians and logicians (via the contradictory principle of comprehension). The axiom of choice (AC) is the principle that,

4. Cauchy's reasoning was clearly nonconstructive, or "purely existential" as we have been saying. In order to show that the polynomial must have one root, Cauchy studied the absolute value of the polynomial, which has a global minimum σ . This global minimum is impredicatively defined. Cauchy assumed that it was positive, and from this he derived a contradiction.

given any infinite family of disjoint nonempty sets, there is a set, known as a *choice set*, that contains exactly one element from each set in the family. The problem with this, said the critics, is that it merely stipulates the existence of the choice set and does not give a defining property for it. Indeed, when it is possible to characterize the choice set explicitly, then the use of AC is avoidable! But in the case of Zermelo's well-ordering theorem it is essential to employ AC. The required well-ordering of \mathbb{R} "exists" in the ideal sense of Cantor, Dedekind, and Hilbert, but it seemed clear that it was completely out of reach from any constructivist perspective.

Thus, the axiom of choice exacerbated obscurities in previous conceptions of set theory, forcing mathematicians to introduce much-needed clarifications. On the one hand, AC was nothing but an explicit statement of previous views about *arbitrary* subsets, and yet, on the other, it obviously clashed with strongly held views about the need to explicitly define infinite sets by properties. The stage was set for deep debate. The discussions about this particular topic contributed more than anything else to a clarification of the existential implications of modern mathematical methods. It is instructive to know that BOREL [VI.70], Baire, and LEBESGUE [VI.72], who became critics, had all relied on AC in less obvious ways in order to prove theorems of analysis. Not by chance, the axiom was suggested to Zermelo by an analyst, Erhard Schmidt, who was a student of Hilbert.⁵

After the publication of Zermelo's proof, an intense debate developed throughout Europe. Zermelo was spurred on to work out the foundations of set theory in an attempt to show that his proof could be developed within an unexceptionable axiom system. The outcome was his famous AXIOM SYSTEM [IV.1 §3], a masterpiece that emerged from careful analysis of set theory as it was historically given in the contributions of Cantor and Dedekind and in Zermelo's own theorem. With some additions due to Fraenkel and VON NEUMANN [VI.91] (the axioms of replacement and regularity) and the major innovation proposed by Weyl and SKOLEM [VI.81] (to formulate it within first-order logic, i.e., quantifying over individuals, the sets, but not over their properties), the axiom system became in the 1920s the one that we now know.

The ZFC system (this stands for "Zermelo-Fraenkel with choice") codifies the key traits of modern mathematical methodology, offering a satisfactory framework for the development of mathematical theories and the conduct of proofs. In particular, it includes strong existence principles, allows impredicative definitions and arbitrary functions, warrants purely existential proofs, and makes it possible to define the main mathematical structures. It thus exhibits all the tendencies (i)–(iv) mentioned in section 1. Zermelo's own work was completely in line with Hilbert's informal axiomatizations of about 1900, and he did not forget to promise a proof of consistency. Axiomatic set theory, whether in the Zermelo-Fraenkel presentation or the von Neumann-Bernays-Gödel version, is the system that most mathematicians regard as the working foundation for their discipline.

As of 1910, the contrast between Russell's type theory and Zermelo's set theory was strong. The former system was developed within formal logic, and its point of departure (albeit later compromised for pragmatic reasons) was in line with predicativism; in order to derive mathematics, the system needed the existential assumptions of infinity and choice, but these were rhetorically treated as tentative hypotheses rather than outright axioms. The latter system was presented informally, adopted the impredicative standpoint wholeheartedly, and asserted as axioms strong existential assumptions that were sufficient to derive all of classical mathematics and Cantor's theory of the higher infinite. In the 1920s the separation diminished greatly, especially with respect to the first two traits just indicated. Zermelo's system was perfected and formulated within the language of modern formal logic. And the Russellians adopted simple type theory, thus accepting the impredicative and "existential" methodology of modern mathematics. This is often given the (potentially confusing) term "Platonism": the objects that the theory refers to are treated *as if* they were independent of what the mathematician can actually and explicitly define.

Meanwhile, back in the first decade of the twentieth century, a young mathematician in the Netherlands was beginning to find his way toward a philosophically colored version of constructivism. Brouwer presented his strikingly peculiar metaphysical and ethical views in 1905, and started to elaborate a corresponding foundation for mathematics in his thesis of 1907. His philosophy of "intuitionism" derived from the old metaphysical view that individual consciousness is the one and

5. One may still gain much insight by reading the letters exchanged by the French analysts in 1905 (see Moore 1982; Ewald 1996) and Zermelo's clever arguments in his second 1908 proof of well-ordering (van Heijenoort 1967).

only source of knowledge. This philosophy is perhaps of little interest in itself, so we shall concentrate here on Brouwer's constructivistic principles. In the years around 1910, Brouwer became a renowned mathematician, with crucial contributions to topology such as his FIXED-POINT THEOREM [V.13]. By the end of World War I, he started to publish detailed elaborations of his foundational ideas, helping to create the famous "crisis," to which we now turn. He was also successful in establishing the customary (but misleading) distinction between formalism and intuitionism.

3 The Crisis in a Strict Sense

In 1921, the *Mathematische Zeitschrift* published a paper by Weyl in which the famous mathematician, who was a disciple of Hilbert, openly espoused intuitionism and diagnosed a "crisis in the foundations" of mathematics. The crisis pointed toward a "dissolution" of the old state of analysis, by means of Brouwer's "revolution." Weyl's paper was meant as a propaganda pamphlet to rouse the sleepers, and it certainly did. Hilbert answered in the same year, accusing Brouwer and Weyl of attempting a "putsch" aimed at establishing "dictatorship à la Kronecker" (see the relevant papers in Mancosu (1998) and van Heijenoort (1967)). The foundational debate shifted dramatically toward the battle between Hilbert's attempts to justify "classical" mathematics and Brouwer's developing reconstruction of a much-reformed intuitionistic mathematics.

Why was Brouwer "revolutionary"? Up to 1920 the key foundational issues had been the acceptability of the real numbers and, more fundamentally, of the impredicativity and strong existential assumptions of set theory, which supported the higher infinite and the unrestricted use of existential proofs. Set theory and, by implication, classical analysis had been criticized for their reliance on impredicative definitions and for their strong existential assumptions (in particular, the axiom of choice, of which extensive use was made by SIERPIŃSKI [VI.77] in 1918). Thus, the debate in the first two decades of the twentieth century was mainly about which principles to accept when it came to defining and establishing the existence of sets and subsets. A key question was, can one make rigorous the vague idea behind talk of "arbitrary subsets"? The most coherent reactions had been Zermelo's axiomatization of set theory and Weyl's predicative system in *Das Kontinuum*. (The *Principia Mathematica* of Whitehead and Russell was an unsuccessful compromise between predicativism and classical mathematics.)

Brouwer, however, brought new and even more basic questions to the fore. No one had questioned the traditional ways of reasoning about the natural numbers: classical logic, in particular the use of quantifiers and the principle of the excluded middle, had been used in this context without hesitation. But Brouwer put forward principled critiques of these assumptions and started developing an alternative theory of analysis that was much more radical than Weyl's. In doing so, he came upon a new theory of the continuum, which finally enticed Weyl and made him announce the coming of a new age.

3.1 Intuitionism

Brouwer began the systematic development of his views with two papers on "intuitionistic set theory," written in German and published in 1918 and 1919 by the *Verhandelungen* of the Dutch Academy of Sciences. These contributions were part of what he regarded as the "Second Act" of intuitionism. The "First Act" (from 1907) had been his emphasis on the intuitive foundations of mathematics. Already Klein and Poincaré had insisted that intuition has an inescapable role to play in mathematical knowledge: as important as logic is in proofs and in the development of mathematical theory, mathematics cannot be reduced to pure logic; theories and proofs are of course organized logically, but their basic principles (axioms) are grounded in intuition. But Brouwer went beyond them and insisted on the absolute independence of mathematics from language and logic.

From 1907, Brouwer rejected the principle of the excluded middle (PEM), which he regarded as equivalent to Hilbert's conviction that all mathematical problems are solvable. PEM is the logical principle that the statement $p \vee \neg p$ (that is, either p or not p) must always be true, whatever the proposition p may be. (For example, it follows from PEM that either the decimal expansion of π contains infinitely many sevens or it contains only finitely many sevens, even though we do not have a proof of which.) Brouwer held that our customary logical principles were abstracted from the way we dealt with subsets of a finite set, and that it was wrong to apply them to infinite sets as well. After World War I he started the systematic reconstruction of mathematics.

The intuitionist position is that one can only state " p or q " when one can give either a constructive proof of p or a constructive proof of q . This standpoint has the consequence that proofs by contradiction (*reductio ad*

PUP: editors think that italics are more appropriate than quotes here, and are more in keeping with usage elsewhere in the volume.

absurdum) are not valid. Consider Hilbert's first proof of his basis theorem (section 1), achieved by *reductio*: he showed that one can derive a contradiction from the assumption that the basis is infinite, and from this he concluded that the basis is finite. The logic behind this procedure is that we start from a concrete instance of PEM, $p \vee \neg p$, show that $\neg p$ is untenable, and conclude that p must be true. But constructive mathematics asks for *explicit* procedures for constructing each object that is assumed to exist, and explicit constructions behind any mathematical statement. Similarly, we have mentioned before (section 2.1) Cauchy's proof of the fundamental theorem of algebra, as well as many proofs in real analysis that invoke the least upper bound. All of these proofs are invalid for a constructivist, and several mathematicians have tried to save the theorems by finding constructivist proofs for them. For instance, both Weyl and Kneser worked on constructivist proofs for the fundamental theorem of algebra.

It is easy to give instances of the use of PEM that a constructivist will not accept: one just has to apply it to any unsolved mathematical problem. For example, Catalan's constant is the number

$$K = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)^2}.$$

It is not known whether K is transcendental, so if p is the statement "Catalan's constant is transcendental," then a constructivist will not accept that p is either true or false.

This may seem odd, or even obviously wrong, until one realizes that constructivists have a different view about what truth *is*. For a constructivist, to say that a proposition is true simply *means* that we can prove it in accordance with the stringent methods that we are discussing; to say that it is false *means* that we can actually exhibit a counterexample to it. Since there is no reason to suppose that every existence statement has either a strict constructivist proof or an explicit counterexample, there is no reason to believe PEM (with this notion of truth). Thus, in order to establish the existence of a natural number with a certain property, a proof by *reductio ad absurdum* is not enough. Existence must be shown by explicit determination or construction if you want to persuade a constructivist.

Notice also how this viewpoint implies that mathematics is not timeless or ahistorical. It was only in 1882 that Lindemann proved that π is a TRANSCENDENTAL NUMBER [III.43]. Since that date, it has been possible to assign a truth value to statements that were

neither true nor false before, according to intuitionists. This may seem paradoxical, but it was just right for Brouwer, since in his view mathematical objects are mental constructions and he rejected as "metaphysics" the assumption that they have an independent existence.

In 1918, Brouwer replaced the sets of Cantor and Zermelo by constructive counterparts, which he would later call "spreads" and "species." A *species* is basically a set that has been defined by a characteristic property, but with the proviso that *every* element has been previously and independently defined by an explicit construction. In particular, the definition of any given species will be strictly predicative.

The concept of a *spread* is particularly characteristic of intuitionism, and it forms the basis for Brouwer's definition of the continuum. It is an attempt to avoid idealization and do justice to the temporal nature of mathematical constructions. Suppose, for example, that we wish to define a sequence of rational numbers that gives better and better approximations to the square root of 2. In classical analysis, one conceives of such sequences as existing in their entirety, but Brouwer defined a notion that he called a *choice sequence*, which pays more attention to how they might be produced. One way to produce them is to give a rule, such as the recurrence relation $x_{n+1} = (x_n^2 + 2)/2x_n$ (and the initial condition $x_1 = 2$). But another is to make less rigidly determined choices that obey certain constraints: for instance, one might insist that x_n has denominator n and that x_n^2 differs from 2 by at most $100/n$, which does not determine x_n uniquely but does ensure that the sequence produces better and better approximations to $\sqrt{2}$.

A choice sequence is therefore not required to be completely specified from the outset, and it can involve choices that are freely made by the mathematician at different moments in time. Both these features make choice sequences very different from the sequences of classical analysis: it has been said that intuitionist mathematics is "mathematics in the making." By contrast, classical mathematics is marked by a kind of timeless objectivity, since its objects are assumed to be fully determined in themselves and independent of the thinking processes of mathematicians.

A spread has choice sequences as its elements—it is something like a law that regulates how the sequences are constructed.⁶ For instance, one could take a spread

6. More precisely, a spread is defined by means of two laws; see Heyting (1956), or more recently van Atten (2003), for further details

that consisted of all choice sequences that began in some particular way, and such a spread would represent a segment—in general, spreads do not represent isolated elements, but continuous domains. By using spreads whose elements satisfy the Cauchy condition, Brouwer offered a new mathematical conception of the *continuum*: rather than being made up of points (or real numbers) with some previous Platonic existence, it was more genuinely “continuous.” Interestingly, this view is reminiscent of Aristotle, who, twenty-three centuries earlier, had emphasized the priority of the continuum and rejected the idea that an extended continuum can be made up of unextended points.

The next stage in Brouwer’s redevelopment of analysis was to analyze the idea of a function. Brouwer defined a function to be an assignment of values to the elements of a spread. However, because of the nature of spreads, this assignment had to be wholly dependent on an initial segment of the choice sequence in order to be constructively admissible. This threw up a big surprise: all functions that are everywhere defined are continuous (and even uniformly continuous). What, you might wonder, about the function f where $f(x) = 0$ when $x < 0$ and $f(x) = 1$ when $x \geq 0$? For Brouwer, this is not a well-defined function, and the underlying reason for this is that one can determine spreads for which we do not know (and may never know) whether they are positive, zero, or negative. For instance, one could let x_n be 1 if all the even numbers between 4 and $2n$ are sums of two primes, and -1 otherwise.

The rejection of PEM has the effect that intuitionistic negation differs in meaning from classical negation. Thus, intuitionistic arithmetic is also different from classical arithmetic. Nevertheless, in 1933 Gödel and Gentzen were able to show that the DEDEKIND-PEANO AXIOMS [III.69] of arithmetic are consistent *relative to* formalized intuitionistic arithmetic. (That is, they were able to establish a correspondence between the sentences of both formal systems, such that a contradiction in classical arithmetic yields a contradiction in its intuitionistic counterpart; thus, if the latter is consistent, the former must be as well.) This was a small triumph for the Hilbertians, though corresponding proofs for systems of analysis or set theory have never been found.

on this and other points. One can picture a spread as a subtree of the universal tree of natural numbers (consisting of all finite sequences of natural numbers), together with an assignment of previously available mathematical objects to the nodes. One law of the spread determines nodes in the tree, the other maps them to objects.

Initially there had been hopes that the development of intuitionism would end in a simple and elegant presentation of pure mathematics. However, as Brouwer’s reconstruction developed in the 1920s, it became more and more clear that intuitionistic analysis was extremely complicated and foreign. Brouwer was not worried, for, as he would say in 1933, “the spheres of truth are less transparent than those of illusion.” But Weyl, although convinced that Brouwer had delineated the domain of mathematical intuition in a completely satisfactory way, remarked in 1925: “the mathematician watches with pain the largest part of his towering theories dissolve into mist before his eyes.” Weyl seems to have abandoned intuitionism shortly thereafter. Fortunately, there was an alternative approach that suggested another way of rehabilitating classical mathematics.

3.2 Hilbert’s Program

This alternative approach was, of course, Hilbert’s program, which promised, in the memorable phrasing of 1928, “to eliminate from the world once and for all the skeptical doubts” as to the acceptability of the classical theories of mathematics. The new perspective, which he started to develop in 1904, relied heavily on formal logic and a combinatorial study of the formulas that are provable from given formulas (the axioms). With modern logic, proofs are turned into formal computations that can be checked mechanically, so that the process is purely constructivist.

In the light of our previous discussion (section 1), it is interesting that the new project was to employ Kroneckerian means for a justification of modern, anti-Kroneckerian methodology. Hilbert’s aim was to show that it is impossible to prove a contradictory formula from the axioms. Once this had been shown combinatorially or constructively (or, as Hilbert also said, finitarily), the argument can be regarded as a justification of the axiom system—even if we read the axioms as talking about non-Kroneckerian objects like the real numbers or transfinite sets.

Still, Hilbert’s ideas at the time were marred by a deficient understanding of logical theory.⁷ It was only in 1917–18 that Hilbert returned to this topic, now with a refined understanding of logical theory and a greater awareness of the considerable technical difficulties of

7. The logic he presented in 1905 lagged far behind Frege’s system of 1879 or Peano’s of the 1890s. We do not enter into the development of logical theory in this period (see, for example, Moore 1998).

his project. Other mathematicians played very significant parts in promoting this better understanding. By 1921, helped by his assistant Bernays, Hilbert had arrived at a very refined conception of the formalization of mathematics, and had perceived the need for a deeper and more careful probing into the logical structure of mathematical proofs and theories. His program was first clearly formulated in a talk at Leipzig late in 1922.

Here we will describe the mature form of Hilbert's program, as it was presented for instance in the 1925 paper "On the infinite" (see van Heijenoort 1967). The main goal was to establish, by means of syntactic consistency proofs, the logical acceptability of the principles and modes of inference of modern mathematics. Axiomatics, logic, and formalization made it possible to study mathematical theories from a purely mathematical standpoint (hence the name *metamathematics*), and Hilbert hoped to establish the consistency of the theories by employing very weak means. In particular, Hilbert hoped to answer all of the criticisms of Weyl and Brouwer, and thereby justify set theory, the classical theory of real numbers, classical analysis, and of course classical logic with its PEM (the basis for indirect proofs by *reductio ad absurdum*).

The whole point of Hilbert's approach was to make mathematical theories fully precise, so that it would become possible to obtain precise results about their properties. The following steps are indispensable for the completion of such a program.

- (i) Finding suitable axioms and primitive concepts for a mathematical theory T , such as that of the real numbers.
- (ii) Finding axioms and inference rules for classical logic, which makes the passage from given propositions to new propositions a purely syntactic, formal procedure.
- (iii) Formalizing T by means of the formal logical calculus, so that propositions of T are just strings of symbols, and proofs are sequences of such strings that obey the formal rules of inference.
- (iv) A *finitary* study of the formalized proofs of T that shows that it is impossible for a string of symbols that expresses a contradiction to be the last line of a proof.

In fact, steps (ii) and (iii) can be solved with rather simple systems formalized in first-order logic, like those studied in any introduction to mathematical logic, such

as Dedekind–Peano arithmetic or Zermelo–Fraenkel set theory. It turns out that first-order logic is enough for codifying mathematical proofs, but, interestingly, this realization came rather late—after GÖDEL'S THEOREMS [V.18].

Hilbert's main insight was that, when theories are formalized, any proof becomes a *finite* combinatorial object: it is just an array of strings of symbols complying with the formal rules of the system. As Bernays said, this was like "projecting" the deductive structure of a theory T into the number-theoretic domain, and it became possible to express in this domain the consistency of T . These realizations raised hopes that a finitary study of formalized proofs would suffice to establish the consistency of the theory, that is, to prove the sentence expressing the consistency of T . But this hope, not warranted by the previous insights, turned out to be wrong.⁸

Also, a crucial presupposition of the program was that not only the logical calculus but also each of the axiomatic systems would be *complete*. Roughly speaking, this means that they would be sufficiently powerful to allow the derivation of all the relevant results.⁹ This assumption turned out to be wrong for systems that contain (primitive recursive) arithmetic, as Gödel showed.

It remains to say a bit more about what Hilbert meant by *finitism* (for details, see Tait 1981). This is one of the points in which his program of the 1920s adopted to some extent the principles of intuitionists such as Poincaré and Brouwer and deviated strongly from the ideas Hilbert himself had considered in 1900. The key idea is that, contrary to the views of logicians like Frege and Dedekind, logic and pure thought require something that is given "intuitively" in our immediate experience: the signs and formulas.

In 1905, Poincaré had put forth the view that a formal consistency proof for arithmetic would be circular, as such a demonstration would have to proceed by induction on the length of formulas and proofs, and thus would rely on the same axiom of induction that it was supposed to establish. Hilbert replied in the 1920s that the form of induction required at the metamathematical level is much weaker than full arithmetical induction, and that this weak form is grounded on the

8. For further details, see, for example, Sieg (1999).

9. The notion of "relevant result" should of course be made precise: doing so leads to the notion either of syntactic completeness or of semantic completeness.

finitary consideration of signs that he took to be intuitively given. Finitary mathematics was not in need of any further justification or reduction.

Hilbert's program proceeded gradually by studying weak theories at first and proceeding to progressively stronger ones. The *metatheory* of a formal system studies properties such as consistency, completeness, and some others ("completeness" in the logical sense means that all true or *valid* formulas that can be represented in the calculus are formally deducible in it). Propositional logic was quickly proved to be consistent and complete. First-order logic, also known as *predicate logic*, was proved complete by Gödel in his dissertation of 1929. For all of the 1920s, the attention of Hilbert and coworkers was set on elementary arithmetic and its subsystems; once this had been settled, the project was to move on to the much more difficult, but crucial, cases of the theory of real numbers and set theory. Ackermann and von Neumann were able to establish consistency results for certain subsystems of arithmetic, but between 1928 and 1930 Hilbert was convinced that the consistency of arithmetic had already been established. Then came the severe blow of Gödel's incompleteness results (see section 4).

The name "formalism," as a description of this program, came from the fact that Hilbert's *method* consisted in formalizing each mathematical theory, and formally studying its proof structure. However, this name is rather one-sided and even confusing, especially because it is usually contrasted with intuitionism, a full-blown *philosophy* of mathematics. Like most mathematicians, Hilbert never viewed mathematics as a mere game played with formulas. Indeed, he often emphasized the meaningfulness of (informal) mathematical statements and the depth of conceptual content expressed in them.¹⁰

3.3 Personal Disputes

The crisis was unfolding not just at an intellectual level but also at a personal level. One should perhaps tell this story as a tragedy, in which the personalities of the main figures and the successive events made the final result quite inescapable.

Hilbert and Brouwer were very different personalities, though they were both extremely willful and clever men. Brouwer's worldview was idealistic and tended

to solipsism. He had an artistic temperament and an eccentric private life. He despised the modern world, looking to the inner life of the self as the only way out (at least in principle, though not always in practice). He preferred to work in isolation, although he had good friends in the mathematical community, especially in the international group of topologists that gathered around him. Hilbert was typically modernist in his views and attitudes; full of optimism and rationalism, he was ready to lead his university, his country, and the international community into a new world. He was very much in favor of collaboration, and felt happy to join Klein's schemes for institutional development and power.

As a consequence of World War I, Germans in the early 1920s were not allowed to attend the International Congresses of Mathematicians. When the opportunity finally arose in 1928, Hilbert was eager to seize on it, but Brouwer was furious because of restrictions that were still imposed on the German delegation and sent a circular letter in order to convince other mathematicians. Their viewpoints were widely known and led to a clash between the two men. On another level, Hilbert had made important concessions to his opponents in the 1920s, hoping that he would succeed in his project of finding a consistency proof. Brouwer emphasized these concessions, accusing him of failing to recognize authorship, and demanded new concessions.¹¹ Hilbert must have felt insulted and perhaps even threatened by a man whom he regarded as perhaps the greatest mathematician of the younger generation.

The last straw came with an episode in 1928. Brouwer had since 1915 been a member of the editorial board of *Mathematische Annalen*, the most prestigious mathematics journal at the time, of which Hilbert had been the main editor since 1902. Ill with "pernicious anemia," and apparently thinking that he was close to the end, Hilbert feared for the future of his journal and decided it was imperative to remove Brouwer from the editorial board. When he wrote to other members of the board explaining his scheme, which he was already carrying out, Einstein replied saying that his proposal was unwise and that he wanted to have nothing to do with it. Other members, however, did not wish to upset the old and admired Hilbert. Finally, a dubious procedure was adopted, where the whole board was dissolved and created anew. Brouwer was greatly disturbed by this

10. This is very explicit, for example, in the lectures of 1919–20 edited by Rowe (1992), and also in the 1930 paper that bears exactly the same title (see *Gesammelte Abhandlungen*, volume 3).

11. See his "Intuitionistic reflections on formalism" of 1928 (in Mancosu 1998).

action, and as a result of it the journal lost Einstein and Carathéodory, who had previously been main editors (see van Dalen 2005).

After that, Brouwer ceased to publish for some years, leaving some book plans unfinished. With his disappearance from the scene, and with the gradual disappearance of previous political turbulences, the feelings of “crisis” began to fade away (see Hesselink 2003). Hilbert did not intervene much in the subsequent debates and foundational developments.

4 Gödel and the Aftermath

It was not only the *Annalen* war that Hilbert won: the mathematical community as a whole continued to work in the style of modern mathematics. And yet his program suffered a profound blow with the publication of Gödel’s famous 1931 article in the *Monatshefte für Mathematik und Physik*. An extremely ingenious development of metamathematical methods—the arithmetization of metamathematics—allowed Gödel to prove that systems like axiomatic set theory or Dedekind–Peano arithmetic are incomplete (see GÖDEL’S THEOREM [V.18]). That is, there exist propositions P formulated strictly in the language of the system such that neither P nor $\neg P$ is formally provable in the system.

This theorem already presented a deep problem for Hilbert’s endeavor, as it shows that formal proof cannot even capture arithmetical truth. But there was more. A close look at Gödel’s arguments made it clear that this first metamathematical proof could itself be formalized, which led to “Gödel’s second theorem”—that it is impossible to establish the consistency of the systems mentioned above by any proof that can be codified *within* them. Gödel’s arithmetization of metamathematics makes it possible to build a sentence, in the language of formal arithmetic, that expresses the consistency of this same formal system. And this sentence turns out to be among those that are unprovable.¹² To express it contrapositively, a finitary formal proof (codifiable in the system of formal arithmetic) of the impossibility of proving $1 = 0$ could be transformed into a contradiction of the system! Thus, if the system is indeed consistent (as most mathematicians are convinced it is), then there is no such finitary proof.

According to what Gödel called at the time “the von Neumann conjecture” (namely, that if there is a

finitary proof of consistency, then it can be formalized and codified within elementary arithmetic), the second theorem implies the failure of Hilbert’s program (see Mancosu (1999, p. 38) and, for more on the reception, Dawson (1997, pp. 68 ff)). One should emphasize that Gödel’s negative results are purely constructivistic and even finitistic, valid for all parties in the foundational debate. They were difficult to digest, but in the end they led to a reestablishment of the basic terms for foundational studies.

Mathematical logic and foundational studies continued to develop brilliantly with Gentzen-style proof theory, with the rise of MODEL THEORY [IV.2], etc.—all of which had their roots in the foundational studies of the first third of the twentieth century. Although the Zermelo–Fraenkel axioms suffice for giving a rigorous foundation to most of today’s mathematics, and have a rather convincing intuitive justification in terms of the “iterative” conception of sets,¹³ there is a general feeling that foundational studies, instead of achieving their ambitious goal, “found themselves attracted into the whirl of mathematical activity, and are now enjoying full voting rights in the mathematical senate.”¹⁴

However, this impression is somewhat superficial. Proof theory has developed, leading to noteworthy reductions of classical theories to systems that can be regarded as constructive. A striking example is that analysis can be formalized in *conservative extensions* of arithmetic: that is, in systems that extend the language of arithmetic while including all theorems of arithmetic, but which are “conservative” in the sense that they have no new consequences in the language of arithmetic. Some parts of analysis can even be developed in conservative extensions of primitive recursive arithmetic (see Feferman 1998). This raises questions about the philosophical bases on which the admissibility of the relevant constructive theories can be founded. But for these systems the question is far less simple than it was for Hilbert’s finitary mathematics; it seems fair to say that no general consensus has yet been reached.

Whatever its roots and justification may be, mathematics is a human activity. This truism is clear from the

12. For further details, see, for example, Smullyan (2001), van Heijenoort (1967), and good introductions to mathematical logic. Both theorems were carefully proved in Hilbert and Bernays (1934/39). Bad expositions and faulty interpretations of Gödel’s results abound.

13. The basic idea is to view the set-theoretic universe as a product of iterating the following operation: one starts with a basic domain V_0 (possibly finite or even equal to \emptyset) and forms all possible *sets of* elements in the domain; this gives a new domain V_1 , and one iterates forming sets of $V_0 \cup V_1$, and so on (to infinity and beyond!). This produces an open-ended set-theoretic universe, masterfully described by Zermelo (1930). On the iterative conception, see, for example, the last papers in Bernacerraf and Putnam (1983).

14. To use the words of Gian-Carlo Rota in an essay of 1973.

subsequent development of our story. The mathematical community refused to abandon “classical” ideas and methods; the constructivist “revolution” was aborted. In spite of its failure, formalism established itself in practice as the avowed ideology of twentieth-century mathematicians. Some have remarked that formalism was less a real faith than a Sunday refuge for those who spent their weekdays working on mathematical objects as something very real. The Platonism of working days was only abandoned, as a BOURBAKI [VI.96] member said, when a ready-made reply was needed to unwelcome philosophical questions concerning mathematical knowledge.

One should note that formalism suited very well the needs of a self-conscious, autonomous community of research mathematicians. It granted them full freedom to choose their topics and to employ modern methods to explore them. However, to reflective mathematical minds it has long been clear that it is not the answer. Epistemological questions about mathematical knowledge have not been “eliminated from the world”; philosophers, historians, cognitive scientists, and others keep looking for more adequate ways of understanding its content and development. Needless to say, this does not threaten the autonomy of mathematical researchers—if autonomy is to be a concern, perhaps we should worry instead about the pressures exerted on us by the market and other powers.

Both (semi-)constructivism and modern mathematics have continued to develop: the contrast between them has simply been consolidated, though in a very unbalanced way, since some 99% of practicing mathematicians are “modern.” (But do statistics matter when it comes to the correct methods for mathematics?) In 1905, commenting on the French debate, HADAMARD [VI.65] wrote that “there are two conceptions of mathematics, two mentalities, in evidence.” It has now come to be recognized that there is value in both approaches: they complement each other and can coexist peacefully. In particular, interest in effective methods, algorithms, and computational mathematics has grown markedly in recent decades—and all of these are closer to the constructivist tradition.

The foundational debate left a rich legacy of ideas and results, key insights and developments, including the formulation of axiomatic set theories and the rise of intuitionism. One of the most important of these developments was the emergence of modern mathematical logic as a refinement of axiomatics, which led to the theories of recursion and computability in around

1936 (see ALGORITHMS [II.4 §3.2]). In the process, our understanding of the characteristics, possibilities, and limitations of formal systems was hugely clarified.

One of the hottest issues throughout the whole debate, and probably its main source, was the question of how to understand the continuum. The reader may recall the contrast between the set-theoretic understanding of the real numbers and Brouwer’s approach, which rejected the idea that the continuum is “built of” points. That this is a labyrinthine question was further established by results on Cantor’s continuum hypothesis (CH), which postulates that the cardinality of the set of real numbers is \aleph_1 , the second transfinite cardinal, or equivalently that every infinite subset of \mathbb{R} must biject with either \mathbb{N} or with \mathbb{R} itself. Gödel proved in 1939 that CH is consistent with axiomatic set theory, but Paul Cohen proved in 1963 that it is independent of its axioms (i.e., Cohen proved that the negation of CH is consistent with axiomatic set theory [IV.1 §5]). The problem is still alive, with a few mathematicians proposing alternative approaches to the continuum and others trying to find new and convincing set-theoretic principles that will settle Cantor’s question (see Woodin 2001).

The foundational debate has also contributed in a definitive way to clarifying the peculiar style and methodology of modern mathematics, especially the so-called Platonism or existential character of modern mathematics (see the classic 1935 paper of Bernays in Benacerraf and Putnam (1983)), by which is meant (here at least) a methodological trait rather than any supposed implications of metaphysical existence. Modern mathematics investigates structures by considering their elements as given independently of human (or mechanical) capabilities of effective definition and construction. This may seem surprising, but perhaps this trait can be explained by broader characteristics of scientific thought and the role played by mathematical structures in the modeling of scientific phenomena.

In the end, the debate made it clear that mathematics and its modern methods are still surrounded by important philosophical problems. When a sizable amount of mathematical knowledge can be taken for granted, theorems can be established and problems can be solved with the certainty and clarity for which mathematics is celebrated. But when it comes to laying out the bare beginnings, philosophical issues cannot be avoided. The reader of these pages may have felt this at several places, especially in the discussion of intuitionism, but also in the basic ideas behind Hilbert’s program, and

of course in the problem of the relationship between formal mathematics and its informal counterpart, a problem that is brought into sharp focus by Gödel's theorems.

I thank Mark van Atten, Jeremy Gray, Paolo Mancosu, José F. Ruiz, Wilfried Sieg, and the editors for their helpful comments on a previous version of this paper.

Further Reading

It is impossible to list here all the relevant articles by Bernays, Brouwer, Cantor, Dedekind, Gödel, Hilbert, Kronecker, von Neumann, Poincaré, Russell, Weyl, Zermelo, etc. The reader can easily find them in the source books by van Heijenoort (1967), Benacerraf and Putnam (1983), Heinzmann (1986), Ewald (1996), and Mancosu (1998).

- van Atten, M. 2003. *On Brouwer*. Belmont, CA: Wadsworth.
- Benacerraf, P., and H. Putnam, eds. 1983. *Philosophy of Mathematics: Selected Readings*. Cambridge: Cambridge University Press.
- van Dalen, D. 1999/2005. *Mystic, Geometer, and Intuitionist: The Life of L. E. J. Brouwer*. Volume I: *The Dawning Revolution*. Volume II: *Hope and Disillusion*. Oxford: Oxford University Press.
- Dawson Jr., J. W. 1997. *Logical Dilemmas: The Life and Work of Kurt Gödel*. Wellesley, MA: A. K. Peters.
- Ewald, W., ed. 1996. *From Kant to Hilbert: A Source Book in the Foundations of Mathematics*, 2 vols. Oxford: Oxford University Press.
- Feferman, S. 1998. *In the Light of Logic*. Oxford: Oxford University Press.
- Ferreirós, J. 1999. *Labyrinth of Thought: A History of Set Theory and Its Role in Modern Mathematics*. Basel: Birkhäuser.
- van Heijenoort, J., ed. 1967. *From Frege to Gödel: A Source Book in Mathematical Logic*. Cambridge, MA: Harvard University Press. (Reprinted, 2002.)
- Heinzmann, G., ed. 1986. *Poincaré, Russell, Zermelo et Peano*. Paris: Vrin.
- Hesseling, D. E. 2003. *Gnomes in the Fog: The Reception of Brouwer's Intuitionism in the 1920s*. Basel: Birkhäuser.
- Heyting, A. 1956. *Intuitionism: An Introduction*. Amsterdam: North-Holland. Third revised edition, 1971.
- Hilbert, D., and P. Bernays. 1934/39. *Grundlagen der Mathematik*, 2 vols. Berlin: Springer.
- Mancosu, P., ed. 1998. *From Hilbert to Brouwer: The Debate on the Foundations of Mathematics in the 1920s*. Oxford: Oxford University Press.
- . 1999. Between Vienna and Berlin: the immediate reception of Gödel's incompleteness theorems. *History and Philosophy of Logic* 20:33–45.
- Mehrtens, H. 1990. *Moderne—Sprache—Mathematik*. Frankfurt: Suhrkamp.

- Moore, G. H. 1982. *Zermelo's Axiom of Choice*. New York: Springer.
- . 1998. Logic, early twentieth century. In *Routledge Encyclopedia of Philosophy*, edited by E. Craig. London: Routledge.
- Rowe, D. 1992. *Natur und mathematisches Erkennen*. Basel: Birkhäuser.
- Sieg, W. 1999. Hilbert's programs: 1917–1922. *The Bulletin of Symbolic Logic* 5:1–44.
- Smullyan, R. 2001. *Gödel's Incompleteness Theorems*. Oxford: Oxford University Press.
- Tait, W. W. 1981. Finitism. *Journal of Philosophy* 78:524–46.
- Weyl, H. 1918. *Das Kontinuum*. Leipzig: Veit.
- Whitehead, N. R., and B. Russell. 1910/13. *Principia Mathematica*. Cambridge: Cambridge University Press. Second edition 1925/27. (Reprinted, 1978.)
- Woodin, W. H. 2001. The continuum hypothesis, I, II. *Notices of the American Mathematical Society* 48:567–76, 681–90.

Part III

Mathematical Concepts

III.1 The Axiom of Choice

Consider the following problem: it is easy to find two irrational numbers a and b such that $a + b$ is rational, or such that ab is rational (in both cases one could take $a = \sqrt{2}$ and $b = -\sqrt{2}$), but is it possible for a^b to be rational? Here is an elegant proof that the answer is yes. Let $x = \sqrt{2}^{\sqrt{2}}$. If x is rational then we have our example. But $x^{\sqrt{2}} = \sqrt{2}^2 = 2$ is rational, so if x is irrational then again we have an example.

Now this argument certainly establishes that it is possible for a and b to be irrational and for a^b to be rational. However, the proof has a very interesting feature: it is *nonconstructive*, in the sense that it does not actually name two irrationals a and b that work. Instead, it tells us that either we can take $a = b = \sqrt{2}$ or we can take $a = \sqrt{2}^{\sqrt{2}}$ and $b = \sqrt{2}$. Not only does it not tell us which of these alternatives will work, it gives us absolutely no clue about how to find out.

Some philosophers and philosophically inclined mathematicians have been troubled by arguments of this kind, but as far as mainstream mathematics goes they are a fully accepted and important style of reasoning. Formally, we have appealed to the “law of the excluded middle.” We have shown that the negation of the statement cannot be true, and deduced that the statement itself must be true. A typical reaction to the proof above is not that it is in any sense invalid, but merely that its nonconstructive nature is rather surprising.

Nevertheless, faced with a nonconstructive proof, it is very natural to ask whether there is a constructive proof. After all, an actual construction may give us more insight into the statement, which is an important point since we prove things not only to be sure they are true but also to gain an idea of *why* they are true. Of course, to ask whether there is a constructive proof is not to suggest that the nonconstructive proof

is invalid, but just that it may be more informative to have a constructive proof.

The *axiom of choice* is one of several rules that we use for building sets out of other sets. Typical examples of such rules are the statement that for any set A we can form the set of all its subsets, and the statement that for any set A and any property p we can form the set of all elements of A that satisfy p (these are usually called the *power-set axiom* and the *axiom of comprehension*, respectively). Roughly speaking, the axiom of choice says that we are allowed to make an arbitrary number of unspecified choices when we wish to form a set.

Like the other axioms, the axiom of choice can seem so natural that one may not even notice that one is using it, and indeed it was applied by many mathematicians before it was first formalized. To get an idea of what it means, let us look at the well-known proof that the union of a countable family of countable sets is countable. The fact that the family is countable allows us to write out the sets in a list A_1, A_2, A_3, \dots , and then the fact that each individual set A_n is countable allows us to write its elements in a list $a_{n1}, a_{n2}, a_{n3}, \dots$. We then finish the proof by finding some systematic way of counting through the elements a_{nm} .

Now in that proof we actually made an infinite number of unspecified choices. We were told that each A_n was countable and then for each A_n we “chose” a listing of the elements of A_n *without specifying the choice we had made*. Moreover, since we are told absolutely nothing about the sets A_n , it is clearly impossible to say how we choose to list them. This remark does not invalidate the proof, but it does show that it is nonconstructive. (Note, however, that if we are actually told what the sets A_n are, then we may well be able to specify listings of their elements and thereby give a constructive proof that the union of those particular sets is countable.)

Here is another example. A GRAPH [III.34] is *bipartite* if its vertices can be split into two classes X and Y in

such a way that no two vertices in the same class are connected by an edge. For example, any even cycle (an even number of points arranged in a circle, with consecutive points joined) is bipartite, while no odd cycle is. Now, is an infinite disjoint union of even cycles bipartite? Of course it is: we just split each of the individual cycles C into two classes X_C and Y_C and then let X be the union of the sets X_C and Y be the union of the sets Y_C . But how do we choose for each cycle C which set to call X_C and which to call Y_C ? Again, we cannot actually specify how we do this, so we are using the axiom of choice (even if we do not explicitly say so).

In general, the axiom of choice states that if we are given a family of nonempty sets X_i , then we may select an element x_i from each one at once. More precisely, it states that if the X_i are nonempty sets, where i ranges over some index set I , then there is a function f defined on I such that $f(i) \in X_i$ for all i . Such a function f is called a *choice function* for the family.

For one set we do not need any separate rule to do this: indeed, the statement that a set X_1 is nonempty is exactly the statement that there exists $x_1 \in X_1$. (More formally, we might say that the function f that takes 1 to x_1 is a choice function for the “family” that consists of the single set X_1 .) For two sets, and indeed for any finite collection of sets, one can prove the existence of a choice function by induction on the number of sets. But for infinitely many sets it turns out that one cannot deduce the existence of a choice function from the other rules for building sets.

Why do people make a fuss about the axiom of choice? The main reason is that if it is used in a proof, then that part of the proof is automatically nonconstructive. This is reflected in the very statement of the axiom. For the other rules that we use, such as “one may take the union of two sets,” the set whose existence is being asserted is uniquely defined by its properties (u is an element of $X \cup Y$ if and only if it is an element of X or of Y or of both). But this is not the case with the axiom of choice: the object whose existence is asserted (a choice function) is not uniquely specified by its properties, and there will typically be many choice functions.

For this reason, the general view in mainstream mathematics is that, although there is nothing wrong with using the axiom of choice, it is a good idea to signal that one has used it, to draw attention to the fact that one’s proof is not constructive.

An example of a statement whose proof involves the axiom of choice is THE BANACH-TARSKI PARADOX

[V.3]. This says that there is a way of dividing up a solid unit sphere into a finite number of subsets and then reassembling these subsets (using rotations, reflections, and translations) to form two solid unit spheres. The proof does not provide an explicit way of defining the subsets.

It is sometimes claimed that the axiom of choice has “undesirable” or “highly counterintuitive” consequences, but in almost all cases a little thought reveals that the consequence under consideration is actually not counterintuitive at all. For example, consider the Banach-Tarski paradox above. Why does it seem strange and paradoxical? It is because we feel that volume has not been preserved. And indeed, this feeling can be converted into a rigorous argument that the subsets used in the decomposition cannot all be sets to which one can meaningfully assign a volume. But that is not a paradox at all: we can say what we mean by the volume of a nice set such as a polyhedron, but there is no reason to suppose that we can give a sensible definition of volume for *all* subsets of the sphere. (The subject called measure theory can be used to give a volume to a very wide class of sets, called the MEASURABLE SETS [III.57], but there is no reason at all to believe that all sets should be measurable, and indeed it can be shown, again by a use of the axiom of choice, that there are sets that are not measurable.)

There are two forms of the axiom of choice that are more often used in daily mathematical life than the basic form we have been discussing. One is the *well-ordering principle*, which states that every set can be well-ordered [III.68]. The other is *Zorn’s lemma*, which states that under certain circumstances “maximal” elements exist. For example, a basis for a vector space is precisely a maximal linearly independent set, and it turns out that Zorn’s lemma applies to the collection of linearly independent sets in a vector space, which shows that every vector space has a basis.

These two statements are called forms of the axiom of choice because they are equivalent to it, in the sense that each one both implies the axiom of choice and may be deduced from it, in the presence of the other rules for building sets. A good way of seeing why these two other forms of the axiom have a nonconstructive feel to them is to spend a few minutes trying to find a well-ordering of the reals, or a basis for the vector space of all sequences of real numbers.

For more about the axiom of choice, and especially about its relationship to the other axioms of formal set theory, see SET THEORY [IV.1].

III.2 The Axiom of Determinacy

Consider the following “infinite game.” Two players, A and B, take turns to name natural numbers, with A going first, say. By doing this, they generate an infinite sequence. A wins the game if this sequence is “eventually periodic,” and B wins if it is not. (An eventually periodic sequence is one like 1, 56, 4, 5, 8, 3, 5, 8, 3, 5, 8, 3, 5, 8, 3, . . . , which settles down after a while to a recurring pattern.) It is not hard to see that B has a winning strategy for this game, since eventually periodic sequences are rather special. However, there is never a point in the game at which B is guaranteed to win, since every finite sequence could be the beginning of an eventually periodic sequence.

More generally, any collection S of infinite sequences of natural numbers gives rise to an infinite game: A's object is now to ensure that the sequence produced is one of the sequences in S , and B's object is to ensure the reverse. The resulting game is called *determined* if one of the two players has a winning strategy. As we have seen, the game is certainly determined when S is the set of eventually periodic sequences, and indeed for just about any set S that one writes down it is easy to see that the corresponding game is determined. Nevertheless, it turns out that there are games that are not determined. (It is an instructive exercise to see where the plausible-seeming argument, “If A does not have a winning strategy, then A cannot force a win, so B must have a winning strategy,” breaks down.)

It is not too hard to construct nondetermined games, but the constructions use THE AXIOM OF CHOICE [III.1]: roughly speaking, one can well-order all possible strategies so that each one has fewer predecessors than there are infinite sequences, and select sequences to belong to S or its complement in a way that stops each strategy in turn from being a winning strategy for either player.

The *axiom of determinacy* states that all games are determined. It contradicts the axiom of choice, but it is a rather interesting axiom when it is added to THE ZERMELO-FRAENKEL AXIOMS [III.101] *without* choice. It turns out, for example, to imply that many sets of reals have surprisingly good properties, such as being Lebesgue measurable. Variants of the axiom of determinacy are closely connected with the theory of large cardinals. For more details, see SET THEORY [IV.1].

Banach Spaces

See NORMED SPACES AND BANACH SPACES [III.64]

III.3 Bayesian Analysis

Suppose you throw a pair of standard dice. The probability that the total is 10 is $\frac{1}{12}$ because there are thirty-six ways the dice can come up, of which three (4 and 6, 5 and 5, and 6 and 4) give 10. If, however, you look at the first die and see that it came up as a 6, then the *conditional probability* that the total is 10, given this information, is $\frac{1}{6}$ (since that is the probability that the other die comes up as a 4).

In general, the *probability of A given B* is defined to be the probability of A and B divided by the probability of B. In symbols, one writes

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \wedge B]}{\mathbb{P}[B]}.$$

From this it follows that $\mathbb{P}[A \wedge B] = \mathbb{P}[A|B] \mathbb{P}[B]$. Now $\mathbb{P}[A \wedge B]$ is the same as $\mathbb{P}[B \wedge A]$. Therefore,

$$\mathbb{P}[A|B] \mathbb{P}[B] = \mathbb{P}[B|A] \mathbb{P}[A],$$

since the left-hand side is $\mathbb{P}[A \wedge B]$ and the right-hand side is $\mathbb{P}[B \wedge A]$. Dividing through by $\mathbb{P}[B]$ we obtain *Bayes's theorem*:

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A] \mathbb{P}[A]}{\mathbb{P}[B]},$$

which expresses the conditional probability of A given B in terms of the conditional probability of B given A.

A fundamental problem in statistics is to analyze random data given by an unknown PROBABILITY DISTRIBUTION [III.73]. Here, Bayes's theorem can make a significant contribution. For example, suppose you are told that an unknown number of unbiased coins between 1 and 10 have been tossed, and that three of them came up heads. Suppose that you wish to guess how many coins there were. Let H_3 stand for the event that three coins came up heads and let C be the number of coins. Then for each n between 1 and 10 it is not hard to calculate the conditional probability $\mathbb{P}[H_3|C = n]$, but we would like to know the reverse, namely $\mathbb{P}[C = n|H_3]$. Bayes's theorem tells us that it is

$$\mathbb{P}[H_3|C = n] \frac{\mathbb{P}[C = n]}{\mathbb{P}[H_3]}.$$

This would tell us the ratios between the various conditional probabilities $\mathbb{P}[C = n|H_3]$ if we knew what the probabilities $\mathbb{P}[C = n]$ were. Typically, one does *not*

know this, but one makes some kind of guess, called a *prior distribution*. For example, one might guess, before knowing that three coins had come up heads, that for each n between 1 and 10 the probability that n coins had been chosen was $\frac{1}{10}$. After this information, one would use the calculation above to revise one's assessment and obtain a *posterior distribution*, in which the probability that $C = n$ would be proportional to $\frac{1}{10} \mathbb{P}[H_3 | C = n]$.

There is more to Bayesian analysis than simply applying Bayes's theorem to replace prior distributions by posterior distributions. In particular, as in the example just given, there is not always an obvious prior distribution to take, and it is a subtle and interesting mathematical problem to devise methods for choosing prior distributions that are "optimal" in different ways. For further discussion, see MATHEMATICS AND MEDICAL STATISTICS [VII.11] and MATHEMATICAL STATISTICS [VII.10].

III.4 Braid Groups

F. E. A. Johnson

Take two parallel planes, each punctured at n points. Label the holes 1 to n in each plane, and run a string from each hole in the first plane to one in the second, in such a way that no two strings go to the same hole. The result is an n -braid. Two different 3-braids, shown in two-dimensional projection in a similar manner to KNOT DIAGRAMS [III.46], are given in figure 1.

As the diagrams suggest, we insist that the strings go from left to right without "doubling back"; so, for example, a knotted string is not allowed.

In describing the "same" braid in different ways, a certain freedom is allowed. Subject to the restrictions that string ends remain fixed and that strings neither break nor pass through each other, strings are allowed to stretch, contract, bend, and otherwise move about in three dimensions. This notion of "sameness" is called *braid isotopy*.

Braids may be composed as follows: arrange a pair of braids end to end to abut in a common (middle) plane, join up the strings, and remove the middle plane. For the braids X and Y in figure 1, the composition XY is given in figure 2.

With this notion of composition, n -braids form a group B_n . In our example, $Y = X^{-1}$, since "pulling all the strings tight" shows that XY is isotopic to the *trivial* braid (figure 3), which acts as the identity.

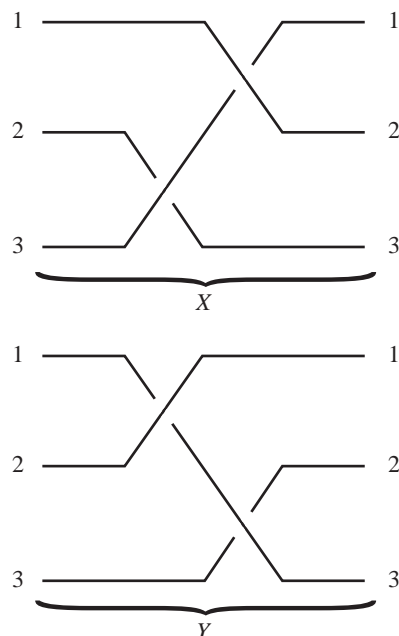


Figure 1 Two 3-braids.

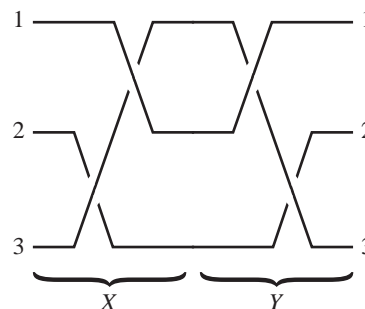


Figure 2 Braid composition.

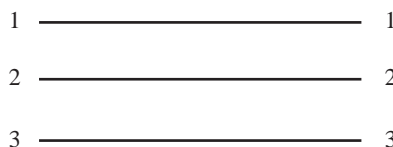
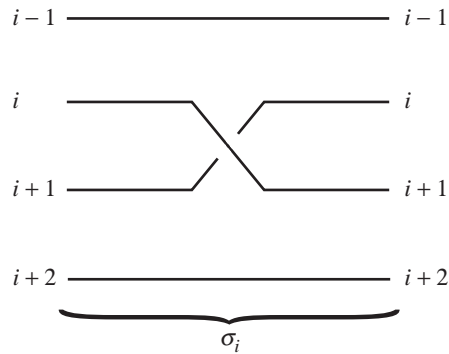


Figure 3 The trivial braid.

As a group, B_n is generated by elements $(\sigma_i)_{1 \leq i \leq n-1}$, where σ_i is formed from the trivial braid by crossing the i th string over the $(i + 1)$ st as in figure 4. The reader may perceive a similarity between the σ_i and the adjacent transpositions that generate the group S_n of

Figure 4 The generator σ_i .

PERMUTATIONS [III.70] of $\{1, \dots, n\}$. Indeed, any braid determines a permutation by the rule

$$i \mapsto \text{right-hand label of } i\text{th string.}$$

Ignoring everything except the behavior at the ends gives a surjective homomorphism $\mathcal{B}_n \rightarrow S_n$, which maps σ_i to the transposition $(i, i+1)$. This is *not* an isomorphism, however, as \mathcal{B}_n is infinite. In fact, σ_i has infinite order, whereas the transposition $(i, i+1)$ squares to the identity. In his celebrated 1925 paper “Theorie der Zöpfe,” ARTIN [VI.86] showed that multiplication in \mathcal{B}_n is completely described by the relations

$$\begin{aligned} \sigma_i \sigma_j &= \sigma_j \sigma_i \quad (|i - j| \geq 2), \\ \sigma_i \sigma_{i+1} \sigma_i &= \sigma_{i+1} \sigma_i \sigma_{i+1}. \end{aligned}$$

These relations have subsequently acquired importance in statistical physics, where they are known as the Yang-Baxter equations.

In groups defined by generators and relations it is usually difficult (there being no method which works uniformly in all cases) to decide whether an arbitrary word in the generators represents the identity element (see GEOMETRIC AND COMBINATORIAL GROUP THEORY [IV.11]). For \mathcal{B}_n , Artin solved this problem geometrically, by “combing the braid.” An alternative algebraic method, due to Garside (1967), also decides when two elements in \mathcal{B}_n are conjugate.

In relation to the decidability of such questions, and in many other respects, braid groups display close affinities with *linear groups*: that is, groups in which all elements behave as if they were invertible $N \times N$ matrices. Although such similarities suggested that it should be possible to prove that braid groups genuinely are linear, the problem of doing so remained unsolved for many years, until in 2001 a proof was eventually found by Bigelow and independently by Krammer.

The groups described here are, strictly speaking, braid groups of the plane, the plane being the object punctured. Other braid groups also occur, often in surprising contexts. The connection with statistical physics has already been mentioned. They arise also in algebraic geometry, when algebraic curves become punctured by discarding exceptional points. Thus, though originating in topology, braids may intervene significantly in areas such as “constructive Galois theory” that seem at first sight to be purely algebraic.

III.5 Buildings

Mark Ronan

The invertible linear transformations on a vector space form a group, called the *general linear group*. If n is the dimension of the vector space and K is the field of scalars, then it is denoted by $GL_n(K)$, and if we pick a basis for the vector space, then each group element can be written as an $n \times n$ matrix whose DETERMINANT [III.15] is nonzero. This group and its subgroups are of great interest in mathematics, and can be studied “geometrically” in the following way. Instead of looking at the vector space V , where of course the origin plays a unique role and is fixed by the group, we use the PROJECTIVE SPACE [I.3 §6.7] associated with V : the points of projective space are the one-dimensional subspaces of V , the lines are the two-dimensional subspaces, the planes are the three-dimensional subspaces, and so on.

Several important subgroups of $GL_n(K)$ can be obtained by imposing constraints on the linear maps (or matrices). For example, $SL_n(K)$ consists of all linear transformations of determinant 1. The group $O(n)$ consists of all linear transformations α of an n -dimensional real inner-product space such that $\langle \alpha v, \alpha w \rangle = \langle v, w \rangle$ for any two vectors v and w (or in matrix terms all real matrices A such that $AA^T = I$); more generally, one can define many similar subgroups of $GL_n(K)$ by taking all linear maps that preserve certain forms, such as bilinear or sesquilinear forms. These subgroups are called *classical groups*. The classical groups are either simple or close to simple (for example, we can often make them simple by quotienting out by the subgroup of scalar matrices). When K is the field of real or complex numbers, the classical groups are *Lie groups*.

Lie groups and their classification are discussed in LIE THEORY [III.50]: the simple Lie groups comprise the classical groups, which fall into one of four families, known as A_n , B_n , C_n , and D_n (where n is a natural

number), along with other types known as E_6 , E_7 , E_8 , F_4 , and G_2 . The subscripts are related to the dimensions of the groups. For example, the groups of type A_n are the groups of invertible linear transformations in $n + 1$ dimensions.

These simple Lie groups have analogues over any field, where they are often referred to as *groups of Lie type*. For example, K can be a finite field, in which case the groups are finite. It turns out that almost all finite simple groups are of Lie type: see THE CLASSIFICATION OF FINITE SIMPLE GROUPS [V.8]. A geometric theory underlying the classical groups had been developed by the first half of the twentieth century. It used projective space and various subgeometries of projective space, which made it possible to provide analogues for the classical groups, but it did not provide analogues for the groups of types E_6 , E_7 , E_8 , F_4 , and G_2 . For this reason, Jacques Tits looked for a geometric theory that would embrace all families, and ended up creating the theory of *buildings*.

The full abstract definition of a building is somewhat complicated, so instead we shall try to give some idea of the concept by looking at the building associated with the groups $GL_n(K)$ and $SL_n(K)$, which are of type A_{n-1} . This building is an *abstract simplicial complex*, which can be thought of as a higher-dimensional analogue of a GRAPH [III.34]. It consists of a collection of points called *vertices*; as in a graph, some pairs of vertices form *edges*; however, it is then possible for triples of vertices to form two-dimensional *faces*, and for sets of k vertices to form $(k - 1)$ -dimensional “simplexes.” (The geometrical meaning of the word “simplex” is a convex hull of a finite set of points in general position: for instance, a three-dimensional simplex is a tetrahedron.) All faces of simplexes must also be included, so for example three vertices cannot form a two-dimensional face unless each pair is joined by an edge.

To form the building of type A_{n-1} , we start by taking all the 1-spaces, 2-spaces, 3-spaces, and so on (corresponding to points, lines, planes, and so on, in projective space), and treat them as “vertices.” The simplexes are formed by all nested sequences of proper subspaces: for example, a 2-space inside a 4-space inside a 5-space will form a “triangle” whose vertices are these three subspaces. The simplexes of maximal dimension have $n - 1$ vertices: a 1-space inside a 2-space inside a 3-space, and so on. These simplexes are called *chambers*.

There are many subspaces, so a building is a huge object. However, buildings have important subgeometries called *apartments*, which in the A_{n-1} case are obtained by taking a basis for the vector space, and then taking all subspaces generated by subsets of this basis. For example, in the A_3 case our vector space is four dimensional, so a basis has four elements; its subsets span four 1-spaces, six 2-spaces, and four 3-spaces. To visualize this apartment it helps to view the four 1-spaces as the vertices of a tetrahedron, the six 2-spaces as the midpoints of its edges, and the four 3-spaces as the midpoints of its faces. The apartment has twenty-four chambers, six for each face of the original tetrahedron, and they form a triangular tiling of the surface of the tetrahedron. This surface is topologically equivalent to a sphere, as are all apartments of this building: such buildings are called *spherical*. The buildings for the groups of Lie type are all spherical, and, just as A_3 is related to the tetrahedron, their apartments are related to the regular and semiregular polyhedra in n dimensions, where n is the subscript in the Lie notation given earlier.

Buildings have the following two noteworthy features. First, any two chambers lie in a common apartment: this is not obvious in the example above but it can be proved using linear algebra. Second, in any building all apartments are isomorphic and any two apartments intersect nicely: more precisely, if A and A' are apartments, then $A \cap A'$ is convex and there is an isomorphism from A to A' that fixes $A \cap A'$. These two features were originally used by Tits in defining buildings.

The theory of spherical buildings does not just give a pleasing geometric basis for the groups of Lie type: it can also be used to construct those of types E_6 , E_7 , E_8 , and F_4 , for an arbitrary field K , without the need for sophisticated machinery such as Lie algebras. Once the building has been constructed (and a construction can be given in a surprisingly simple manner), a theorem of Tits on the existence of automorphisms shows that the groups themselves must exist.

In a spherical building the apartments are tilings of a sphere, but other types of buildings also play significant roles. Of particular importance are *affine buildings*, in which the apartments are tilings of Euclidean space; such buildings arise in a natural way from groups, such as $GL_n(K)$, where K is a p -ADIC FIELD [III.53]. For such fields there are two buildings, one spherical and one affine, but the affine one carries more information and yields the spherical building as a structure “at infinity.” Going beyond affine buildings, there are hyperbolic

PUP: I can confirm that this is correct as written.

PUP: some very minor corrections suggested by author after proofreading proof sent to you and these are included above. The main one, though, is the addition of this paragraph.

buildings, whose apartments are tilings of hyperbolic space; they arise naturally in the study of hyperbolic Kac-Moody groups.

III.6 Calabi-Yau Manifolds

Eric Zaslow

1 Basic Definition

Calabi-Yau manifolds, named after Eugenio Calabi and Shing-Tung Yau, arise in Riemannian geometry and algebraic geometry, and play a prominent role in string theory and mirror symmetry.

In order to explain what they are, we need first to recall the notion of orientability on a real MANIFOLD [I.3 §6.9]. Such a manifold is *orientable* if you can choose coordinate systems at each point in such a way that any two systems $x = (x^1, \dots, x^m)$ and $y = (y^1, \dots, y^m)$ that are defined on overlapping sets give rise to a positive Jacobian: $\det(\partial y^i / \partial x^j) > 0$. The notion of a Calabi-Yau manifold is the natural complex analogue of this. Now the manifold is complex, and for each local coordinate system $z = (z^1, \dots, z^n)$ one has a HOLOMORPHIC FUNCTION [I.3 §5.6] $f(z)$. It is vital that f should be *nonvanishing*: that is, it never takes the value 0. There is also a compatibility condition: if $\tilde{z}(z)$ is another coordinate system, then the corresponding function \tilde{f} is related to f by the equation $f = \tilde{f} \det(\partial \tilde{z}^a / \partial z^b)$. Note that in this definition, if we replace all complex terms by real terms, then we have the notion of a real orientation. So a Calabi-Yau manifold can be thought of informally as a complex manifold with complex orientation.

2 Complex Manifolds and Hermitian Structure

Before we go any further, a few words about complex and Kähler geometry are in order. A complex manifold is a structure that looks locally like \mathbb{C}^n , in the sense that one can find complex coordinates $z = (z^1, \dots, z^n)$ near every point. Moreover, where two coordinate systems z and \tilde{z} overlap, the coordinates \tilde{z}^a are holomorphic when they are regarded as functions of the z^b . Thus the notion of a holomorphic function on a complex manifold makes sense and does not depend on the coordinates used to express the function. In this way, the local geometry of a complex manifold does indeed look like an open set in \mathbb{C}^n , and the tangent space at a point looks like \mathbb{C}^n itself.

On complex vector spaces it is natural to consider Hermitian INNER PRODUCTS [III.37] represented by HERMITIAN MATRICES [III.52 §3]¹ $g_{a\bar{b}}$ with respect to a basis e_a . On complex manifolds, a Hermitian inner product on the tangent spaces is called a “Hermitian metric,” and is represented in a coordinate basis by a Hermitian matrix $g_{a\bar{b}}$, which depends on position.

3 Holonomy, and Calabi-Yau Manifolds in Riemannian Geometry

On a RIEMANNIAN MANIFOLD [I.3 §6.10] one can move a vector along a path so as to keep it of constant length and “always pointing in the same direction.” *Curvature* expresses the fact that the vector you wind up with at the end of the path depends on the path itself. When your path is a closed loop, the vector at the starting point comes back to a new vector at the same point. (A good example to think about is a path on a sphere that goes from the North Pole to the equator, then a quarter of the way around the equator, then back to the North Pole again. When the journey is completed, the “constant” vector that began by pointing south will have been rotated by 90° .) With each loop we associate a matrix operator, called the *holonomy matrix*, which sends the starting vector to the ending vector; the group generated by all of these matrices is called the *holonomy group* of the manifold. Since the length of the vector does not change during the process of keeping it constant along the loop, the holonomy matrices all lie in the orthogonal group of length-preserving matrices, $O(m)$. If the manifold is oriented, then the holonomy group must lie in $SO(m)$, as one can see by transporting an oriented basis of vectors around the loop.

Every complex manifold of (complex) dimension n is also a real manifold of (real) dimension $m = 2n$, which one can think of as coordinatized by the real and imaginary parts of the complex coordinates z^j . Real manifolds that arise in this way have additional structure. For example, the fact that we can multiply complex coordinate directions by $i = \sqrt{-1}$ implies that there must be an operator on the real tangent space that squares to -1 . This operator has eigenvalues $\pm i$, which can be thought of as “holomorphic” and “anti-holomorphic” directions. The Hermitian property states that these directions are orthogonal, and we say that the manifold is *Kähler* if they remain so after

1. The notation $g_{a\bar{b}}$ indicates the conjugate-linear property of a Hermitian inner product.

PUP: I can confirm that bold 1 is OK.

PUP: this ‘the’ has to stay. It shows the reader that there is more than one ‘ z^b ’ and is a pretty common formulation in maths writing.

transport around loops. This means that the holonomy group is a subgroup of $U(n)$ (which itself is a subgroup of $SO(2m)$: complex manifolds always have *real* orientations). There is a nice local characterization of the Kähler property: if $g_{a\bar{b}}$ are the components of the Hermitian metric in some coordinate patch, then there exists a function φ on that patch such that $g_{a\bar{b}} = \partial^2 \varphi / \partial z^a \partial \bar{z}^b$.

Given a complex orientation—that is, the nonmetric definition of a Calabi-Yau manifold given above—a *compatible Kähler structure* leads to a holonomy that lies in $SU(n) \subset U(n)$, the natural analogue of the case of real orientation. This is the metric definition of a Calabi-Yau manifold.

4 The Calabi Conjecture

Calabi conjectured that, for any Kähler manifold of complex dimension n and any complex orientation, there exists a function u and a new Kähler metric \tilde{g} , given in coordinates by

$$\tilde{g}_{a\bar{b}} = g_{a\bar{b}} + \frac{\partial^2 u}{\partial z^a \partial \bar{z}^b},$$

that is compatible with the orientation. In equations, the compatibility condition states that

$$\det \left(g_{a\bar{b}} + \frac{\partial^2 u}{\partial z^a \partial \bar{z}^b} \right) = |f|^2,$$

where f is the holomorphic orientation function discussed above. Thus, the metric notion of a Calabi-Yau manifold amounts to a formidable nonlinear partial differential equation for u . Calabi proved the uniqueness and Yau proved the existence of a solution to this equation. So in fact the metric definition of a Calabi-Yau manifold is uniquely determined by its Kähler structure and its complex orientation.

Yau's theorem establishes that the space of metrics with holonomy group $SU(n)$ on a manifold with complex orientation is in correspondence with the space of inequivalent Kähler structures. The latter space can easily be probed with the techniques of algebraic geometry.

5 Calabi-Yau Manifolds in Physics

Einstein's theory of gravity, general relativity, constructs equations that the metric of a Riemannian space-time manifold must obey (see GENERAL RELATIVITY AND THE EINSTEIN EQUATIONS [IV.17]). The equations involve three symmetric tensors: the metric,

the RICCI CURVATURE [III.80] tensor, and the energy-momentum tensor of matter. A Riemannian manifold whose Ricci tensor vanishes is a solution to these equations when there is no matter, and is a special case of an *Einstein manifold*. A Calabi-Yau manifold with its unique $SU(n)$ -holonomy metric has vanishing Ricci tensor, and is therefore of interest in general relativity.

A fundamental problem in theoretical physics is the incorporation of Einstein's theory into the quantum theory of particles. This enterprise is known as *quantum gravity*, and Calabi-Yau manifolds figure prominently in the leading theory of quantum gravity, STRING THEORY [IV.13 §2].

In string theory, the fundamental objects are one-dimensional “strings.” The motion of the strings through space-time is described by two-dimensional trajectories, known as *worldsheets*, so every point on the worldsheet is labeled by the point in space-time where it sits. In this way, string theory is constructed from a quantum field theory of maps from two-dimensional RIEMANN SURFACES [III.81] to a space-time manifold M . The two-dimensional surface should be given a Riemannian metric, and there is an infinite-dimensional space of such metrics to consider. This means that we must solve quantum gravity in two dimensions—a problem that, like its four-dimensional cousin, is too hard. If, however, it happens that the two-dimensional worldsheet theory is conformal (invariant under local changes of scale), then just a finite-dimensional space of conformally inequivalent metrics remains, and the theory is well-defined.

The Calabi-Yau condition arises from these considerations. The requirement that the two-dimensional theory is conformal, so that the string theory makes good sense, is in essence the requirement that the Ricci tensor of space-time should vanish. Thus, a two-dimensional condition leads to a space-time equation, which turns out to be exactly Einstein's equation without matter. We add to this condition the “phenomenological” criterion that the theory be endowed with “supersymmetry,” which requires the space-time manifold M to be complex. The two conditions together mean that M is a complex manifold with holonomy group $SU(n)$: that is, a Calabi-Yau manifold. By Yau's theorem, the choices of such M can easily be described by algebraic geometric methods.

We remark that there is a kind of distillation of string theory called “topological strings,” which can be given a rigorous mathematical framework. Calabi-Yau manifolds are both symplectic and complex, and this leads

to two versions of topological strings, called A and B, that one can associate with a Calabi–Yau manifold. Mirror symmetry is the remarkable phenomenon that the A version of one Calabi–Yau manifold is related to the B version of an entirely different “mirror partner.” The mathematical consequences of such an equivalence are extremely rich. (See MIRROR SYMMETRY [IV.14] for more details. For other notions related to those discussed in this article, see SYMPLECTIC MANIFOLDS [III.90].)

The Calculus of Variations

See VARIATIONAL METHODS [III.96]

III.7 Cardinals

The cardinality of a set is a measure of how large that set is. More precisely, two sets are said to have the same cardinality if there is a bijection between them. So what do cardinalities look like?

There are finite cardinalities, meaning the cardinalities of finite sets: a set has “cardinality n ” if it has precisely n elements. Then there are COUNTABLE [III.11] infinite sets: these all have the same cardinality (this follows from the definition of “countable”), usually written \aleph_0 . For example, the natural numbers, the integers, and the rationals all have cardinality \aleph_0 . However, the reals are uncountable, and so do not have cardinality \aleph_0 . In fact, their cardinality is denoted by 2^{\aleph_0} .

It turns out that cardinals can be added and multiplied and even raised to powers of other cardinals (so “ 2^{\aleph_0} ” is not an isolated piece of notation). For details, and more explanation, see SET THEORY [IV.1 §2].

III.8 Categories

Eugenia Cheng

When we study GROUPS [I.3 §2.1] or VECTOR SPACES [I.3 §2.3], we pay particular attention to certain classes of maps between them: the important maps between groups are the group HOMOMORPHISMS [I.3 §4.1], and between vector spaces they are the LINEAR MAPS [I.3 §4.2]. What makes these maps important is that they are the functions that “preserve structure”: for example, if ϕ is a homomorphism from a group G to a group H , then it “preserves multiplication,” in the sense that $\phi(g_1 g_2) = \phi(g_1) \phi(g_2)$ for any pair of elements g_1 and g_2 of G . Similarly, linear maps preserve addition and scalar multiplication.

The notion of a structure-preserving map applies far more generally than just to these two examples, and one of the purposes of category theory is to understand the general properties of such maps. For instance, if A , B , and C are mathematical structures of some given type, and f and g are structure-preserving maps from A to B and from B to C , respectively, then their composite $g \circ f$ is a structure-preserving map from A to C . That is, structure-preserving maps can be *composed* (at least if the range of one equals the domain of the other). We also use structure-preserving maps to decide when to regard two structures as “essentially the same”: we call A and B *isomorphic* if there is a structure-preserving map from A to B with an inverse that also preserves structure.

A *category* is a mathematical structure that allows one to discuss properties such as these in the abstract. It consists of a collection of *objects*, together with *morphisms* between those objects. That is, if a and b are two objects in the category, then there is a collection of morphisms between a and b . There is also a notion of *composition* of morphisms: if f is a morphism from a to b and g is a morphism from b to c , then there is a *composite* of f and g , which is a morphism from a to c . This composition must be associative. In addition, for each object a there is an “identity morphism,” which has the property that if you compose it with another morphism f then you get f .

As the earlier discussion suggests, an example of a category is the category of groups. The objects of this category are groups, the morphisms are group homomorphisms, and composition and the identity are defined in the way we are used to. However, it is by no means the case that all categories are like this, as the following examples show.

- (i) We can form a category by taking the natural numbers as its objects, and letting the morphisms from n to m be all the $n \times m$ matrices with real entries. Composition of morphisms is the usual matrix multiplication. We would not normally think of an $n \times m$ matrix as a map from the number n to the number m , but the axioms for a category are nevertheless satisfied.
- (ii) Any set can be turned into a category: the objects are the elements of the set, and a morphism from x to y is the assertion “ $x = y$.” We can also make an ordered set into a category by letting a morphism from x to y be the assertion “ $x \leq y$.” (The “composite” of “ $x \leq y$ ” and “ $y \leq z$ ” is “ $x \leq z$.”)

- (iii) Any group G can be made into a category as follows: you have just one object, and the morphisms from that object to itself are the elements of the group, with the group multiplication defining the composition of two morphisms.
- (iv) There is an obvious category where the objects are TOPOLOGICAL SPACES [III.92] and the morphisms are continuous functions. A less obvious category with the same objects takes as its morphisms not continuous functions but HOMOTOPY CLASSES [IV.10 §2] of continuous functions.

Morphisms are also called *maps*. However, as the above examples illustrate, the maps in a category do not have to be remotely map-like. They are also called *arrows*, partly to emphasize the more abstract nature of a general category, and partly because arrows are often used to represent morphisms pictorially.

The general framework and language of “objects and morphisms” enable us to seek and study structural features that depend only on the “shape” of the category, that is, on its morphisms and the equations they satisfy. The idea is both to make general arguments that are then applicable to all categories possessing particular structural features, and also to be able to make arguments in specific environments without having to go into the details of the structures in question. The use of the former to achieve the latter is sometimes referred to, endearingly or otherwise, as “abstract nonsense.”

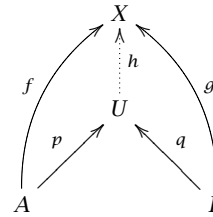
As we mentioned above, the morphisms in a category are generally depicted as arrows, so a morphism f from a to b is depicted as $a \xrightarrow{f} b$ and composition is depicted by concatenating the arrows $a \xrightarrow{f} b \xrightarrow{g} c$. This notation greatly eases complex calculations and gives rise to the so-called *commutative diagrams* that are often associated with category theory; an equality between composites of morphisms such as $g \circ f = t \circ s$ is expressed by asserting that the following diagram *commutes*, that is, that either of the two different paths from a to c yield the same composite:

$$\begin{array}{ccc} a & \xrightarrow{f} & b \\ s \downarrow & & \downarrow g \\ d & \xrightarrow{t} & c \end{array}$$

Proving that one long string of compositions equals another then becomes a matter of “filling in” the space in between with smaller diagrams that are already known to commute. Furthermore, many important mathematical concepts can be described in terms of commutative

diagrams: some examples are free groups, free rings, free algebras, quotients, products, disjoint unions, function spaces, direct and inverse limits, completion, compactification, and geometric realization.

Let us see how it is done in the case of disjoint unions. We say that a *disjoint union* of sets A and B is another set U equipped with morphisms $A \xrightarrow{p} U$ and $B \xrightarrow{q} U$ such that, given any set X and morphisms $A \xrightarrow{f} X$ and $B \xrightarrow{g} X$, there is a unique morphism $U \xrightarrow{h} X$ that makes the following diagram commute:



Here p and q tell us how A and B inject into the disjoint union. The “such that” part of the definition above is a *universal property*. It expresses the fact that giving a function from the disjoint union to another set is precisely the same as giving a function from each of the individual sets; this completely characterizes a disjoint union (which we regard as defined up to isomorphism). Another viewpoint is that the universal property expresses the fact that a disjoint union is the “most free” way of having two sets map into another set, neither adding any information nor collapsing any information. Universal properties are central to the way category theory describes structures that are somehow “canonical.” (See also the discussion of free groups in GEOMETRIC AND COMBINATORIAL GROUP THEORY [IV.11].)

Another key concept in a category is that of an *isomorphism*. As one might expect, this is defined to be a morphism with a two-sided inverse. Isomorphic objects in a given category are thought of as “the same, as far as this particular category is concerned.” Thus, categories provide a framework in which the most natural way of classifying objects is “up to isomorphism.”

Categories are mathematical structures of a certain kind, and as such they themselves form a category (subject to size restrictions so as to avoid a Russell-type paradox). The morphisms, which are the structure-preserving maps for categories, are called *functors*. In other words, a functor F from a category X to a category Y takes the objects of X to the objects of Y and the morphisms of X to the morphisms of Y in such a way that the identity of a is taken to the identity

of Fa and the composite of f and g is taken to the composite of Ff and Fg . An important example of a functor is the one that takes a topological space S with a “marked point” s to its fundamental group $\pi_1(S, s)$: it is one of the basic theorems of algebraic topology that a continuous map between two topological spaces (that takes marked point to marked point) gives rise to a homomorphism between their fundamental groups.

Furthermore, there is a notion of morphism between functors, called a *natural transformation*, which is analogous to the notion of homotopy between maps of topological spaces. Given continuous maps $F, G : X \rightarrow Y$, a homotopy from F to G gives us, for every point x in X , a path in Y from Fx to Gx ; analogously, given functors $F, G : X \rightarrow Y$, a natural transformation from F to G gives us, for every point x in X , a *morphism* in Y from Fx to Gx . There is also a commuting condition that is analogous to the fact that, in the case of homotopy, a path in X must have its image under F continuously transformed to its image under G without passing over any “holes” in the space Y . This avoidance of holes is expressed in the category case by the commutativity of certain squares in the target category Y , which is known as the “naturality condition.”

One example of a natural transformation encodes the fact that every vector space is *canonically* isomorphic to its double dual; there is a functor from the category of vector spaces to itself that takes each vector space to its double dual, and there is an invertible natural transformation from this functor to the identity functor via the canonical isomorphisms. By contrast, every finite-dimensional vector space is isomorphic to its dual, but not canonically so because the isomorphism involves an arbitrary choice of basis; if we attempt to construct a natural transformation in this case, we find that the naturality condition fails. In the presence of natural transformations, categories actually form a 2-*category*, which is a two-dimensional generalization of a category, with objects, morphisms, and morphisms between morphisms. These last are thought of as two-dimensional morphisms; more generally an n -category has morphisms for each dimension up to n .

Categories and the language of categories are used in a wide variety of other branches of mathematics. Historically, the subject is closely associated with algebraic topology; the notions were first introduced in 1945 by Eilenberg and Mac Lane. Applications followed in algebraic geometry, theoretical computer science, theoretical physics, and logic. Category theory, with its abstract nature and lack of dependency on other fields

of mathematics, can be thought of as “foundational.” In fact, it has been proposed as an alternative candidate for the foundations of mathematics, with the notion of morphism as the basic one from which everything else is built up, instead of the relation of set membership that is used in SET-THEORETIC FOUNDATIONS [IV.1 §4].

Class Field Theory

See FROM QUADRATIC RECIPROCITY TO CLASS FIELD THEORY [V.30]

PUP: what do you think of the style of the cross-reference entries now?

Cohomology

See HOMOLOGY AND COHOMOLOGY [III.39]

III.9 Compactness and Compactification

Terence Tao

In mathematics, it is well-known that the behavior of finite sets and the behavior of infinite sets can be rather different. For instance, each of the following statements is easily seen to be true whenever X is a finite set but false whenever X is an infinite set.

All functions are bounded. If $f : X \rightarrow \mathbb{R}$ is a real-valued function on X , then f must be bounded (i.e., there exists a finite number M such that $|f(x)| \leq M$ for all $x \in X$).

All functions attain a maximum. If $f : X \rightarrow \mathbb{R}$ is a real-valued function on X , then there must exist at least one point $x_0 \in X$ such that $f(x_0) \geq f(x)$ for all $x \in X$.

All sequences have constant subsequences. If $x_1, x_2, x_3, \dots \in X$ is a sequence of points in X , then there must exist a subsequence $x_{n_1}, x_{n_2}, x_{n_3}, \dots$ that is constant. In other words, $x_{n_1} = x_{n_2} = \dots = c$ for some $c \in X$. (This fact is sometimes known as the *infinite pigeonhole principle*.)

T&T: check word spacing here at page makeup.

The first statement—that all functions on a finite set are bounded—can be viewed as a very simple example of a *local-to-global principle*. The hypothesis is an assertion of “local” boundedness: it asserts that $|f(x)|$ is bounded for each point $x \in X$ separately, but with a bound that may depend on x . The conclusion is that of “global” boundedness: that $|f(x)|$ is bounded by a *single* bound M for all $x \in X$.

So far we have viewed the object X only as a set. However, in many areas of mathematics we like to

endow our objects with additional structures, such as a TOPOLOGY [III.92], a METRIC [III.58], or a GROUP STRUCTURE [I.3 §2.1]. When we do this, it turns out that some objects exhibit properties similar to those of finite sets (in particular, they enjoy local-to-global principles), even though as sets they are infinite. In the categories of topological spaces and metric spaces, these “almost-finite” objects are known as *compact spaces*. (Other categories have “almost-finite” objects as well. For example, in the category of groups there is a notion of a *pro-finite group*; for LINEAR OPERATORS [III.52] between NORMED SPACES [III.64] the analogous notion is that of a *compact operator*, which is “almost of finite rank”; and so forth.)

A good example of a compact set is the closed unit interval $X = [0, 1]$. This is an infinite set, so the previous three assertions are all false as stated for X . But if we modify them by inserting topological concepts such as continuity and convergence, then we can restore these assertions for $[0, 1]$ as follows.

All continuous functions are bounded. If $f : X \rightarrow \mathbb{R}$ is a real-valued continuous function on X , then f must be bounded. (This is again a type of local-to-global principle: if a function does not vary too much locally, then it does not vary too much globally.)

All continuous functions attain a maximum. If $f : X \rightarrow \mathbb{R}$ is a real-valued continuous function on X , then there must exist at least one point $x_0 \in X$ such that $f(x_0) \geq f(x)$ for all $x \in X$.

All sequences have convergent subsequences. If $x_1, x_2, x_3, \dots \in X$ is a sequence of points in X , then there must exist a subsequence $x_{n_1}, x_{n_2}, x_{n_3}, \dots$ that converges to some limit $c \in X$. (This statement is known as the *Bolzano-Weierstrass theorem*.)

To these assertions we can add a fourth (which, like the others, has a rather trivial analogue for finite sets).

All open covers have finite subcovers. If \mathcal{V} is a collection of open sets and the union of all these open sets contains X (in which case \mathcal{V} is called an *open cover* of X), then there must exist a finite subcollection $V_{n_1}, V_{n_2}, \dots, V_{n_k}$ of sets in \mathcal{V} that still covers X .

All four of these topological statements are false for sets such as the open unit interval $(0, 1)$ or the real line \mathbb{R} , as one can easily check by constructing simple counterexamples. The *Heine-Borel theorem* asserts that when X is a subset of a Euclidean space \mathbb{R}^n , the above

statements are all true when X is topologically closed and bounded, and all false otherwise.

The above four assertions are closely related to each other. For instance, if you know that all sequences in X contain convergent subsequences, then you can quickly deduce that all continuous functions have a maximum. This is done by first constructing a *maximizing sequence*—a sequence of points x_n in X such that $f(x_n)$ approaches the maximal value of f (or, more precisely, its supremum)—and then investigating a convergent subsequence of that sequence. In fact, given some fairly mild assumptions on the space X (e.g., that X is a metric space), one can deduce any of these four statements from any of the others.

To oversimplify a little, we say that a topological space X is *compact* if one (and hence all) of the above four assertions holds for X . Because the four assertions are not quite equivalent in general, the formal definition of compactness uses only the fourth version: that every open cover has a finite subcover. There are other notions of compactness, such as *sequential compactness*, for example, which is based on the third version, but the distinctions between these notions are technical and we shall gloss over them here.

Compactness is a powerful property of spaces, and it is used in many ways in many different areas of mathematics. One is via appeal to local-to-global principles: one establishes local control on a function, or on some other quantity, and then uses compactness to boost the local control to global control. Another is to locate maxima or minima of a function, which is particularly useful in the CALCULUS OF VARIATIONS [III.96]. A third is to partially recover the notion of a limit when dealing with nonconvergent sequences, by accepting the need to pass to a subsequence of the original sequence. (However, different subsequences may converge to different limits; compactness guarantees the existence of a limit point, but not its uniqueness.) Compactness of one object also tends to beget compactness of other objects; for instance, the image of a compact set under a continuous map is still compact, and the product of finitely many or even infinitely many compact sets continues to be compact. This last result is known as *Tychonoff's theorem*.

Of course, many spaces of interest are not compact. An obvious example is the real line \mathbb{R} , which is not compact, because it contains sequences such as $1, 2, 3, \dots$ that are “trying to escape” the real line and that do not leave behind any convergent subsequences. However, one can often recover compactness by adding a few

more points to the space: this process is known as *compactification*. For instance, one can compactify the real line by adding one point at each end: we call the added points $+\infty$ and $-\infty$. The resulting object, known as the *extended real line* $[-\infty, +\infty]$, can be given a topology in a natural way, which basically defines what it means to converge to $+\infty$ or to $-\infty$. The extended real line is compact: any sequence x_n of extended real numbers will have a subsequence that either converges to $+\infty$, converges to $-\infty$, or converges to a finite number. Thus, by using this compactification of the real line, we can generalize the notion of a limit to one that no longer has to be a real number. While there are some drawbacks to dealing with extended reals instead of ordinary reals (for instance, one can always add two real numbers together, but the sum of $+\infty$ and $-\infty$ is undefined), the ability to take limits of what would otherwise be divergent sequences can be very useful, particularly in the theory of infinite series and improper integrals.

It turns out that a single noncompact space can have many different compactifications. For instance, by the device of *stereographic projection*, one can topologically identify the real line with a circle that has a single point removed. (For example, if one maps the real number x to the point $(x/(1+x^2), x^2/(1+x^2))$, then \mathbb{R} maps to the circle of radius $\frac{1}{2}$ and center $(0, \frac{1}{2})$, with the north pole $(0, 1)$ removed.) If we then insert the missing point, we obtain the *one-point compactification* $\mathbb{R} \cup \{\infty\}$ of the real line. More generally, any reasonable topological space (e.g., a locally compact Hausdorff space) has a number of compactifications, ranging from the one-point compactification $X \cup \{\infty\}$, which is the “minimal” compactification as it adds only one point, to the *Stone-Ćech compactification* βX , which is the “maximal” compactification, and adds an enormous number of points. The Stone-Ćech compactification $\beta\mathbb{N}$ of the natural numbers \mathbb{N} is the space of *ultrafilters*, which are very useful tools in the more infinitary parts of mathematics.

One can use compactifications to distinguish between different types of divergence in a space. For instance, the extended real line $[-\infty, +\infty]$ distinguishes between divergence to $+\infty$ and divergence to $-\infty$. In a similar spirit, by using compactifications of the plane \mathbb{R}^2 such as the *PROJECTIVE PLANE* [I.3 §6.7], one can distinguish a sequence that diverges along (or near) the x -axis from a sequence that diverges along (or near) the y -axis. Such compactifications arise naturally in situations in which sequences that diverge in different ways exhibit markedly different behavior.

Another use of compactifications is to allow one to rigorously view one type of mathematical object as a limit of others. For instance, one can view a straight line in the plane as the limit of increasingly large circles by describing a suitable compactification of the space of circles that includes lines. This perspective allows us to deduce certain theorems about lines from analogous theorems about circles, and conversely to deduce certain theorems about very large circles from theorems about lines. In a rather different area of mathematics, the Dirac delta function is not, strictly speaking, a function, but it exists in certain (local) compactifications of spaces of functions, such as spaces of MEASURES [III.57] or DISTRIBUTIONS [III.18]. Thus, one can view the Dirac delta function as a limit of classical functions, and this can be very useful for manipulating it. One can also use compactifications to view the continuous as the limit of the discrete: for instance, it is possible to compactify the sequence $\mathbb{Z}/2\mathbb{Z}, \mathbb{Z}/3\mathbb{Z}, \mathbb{Z}/4\mathbb{Z}, \dots$ of cyclic groups in such a way that their limit is the circle group $\mathbb{T} = \mathbb{R}/\mathbb{Z}$. These simple examples can be generalized to much more sophisticated examples of compactifications, which have many applications in geometry, analysis, and algebra.

III.10 Computational Complexity Classes

One of the basic challenges of theoretical computer science is to determine what computational resources are necessary in order to perform a given computational task. The most basic resource is *time*, or equivalently (given the hardware) the number of steps needed to implement the most efficient algorithm that will actually carry out the task. Especially important is how this time scales up with the size of the input for the task: for instance, how much longer does it take to factorize an integer with $2n$ digits than an integer with n digits? Another resource connected with the feasibility of a computation is *memory*: one can ask how much storage space a computer will need in order to implement an algorithm, and how this can be minimized. A *complexity class* is a set of computational problems that can be performed with certain restrictions on the resources allowed. For instance, the complexity class P consists of all problems that can be performed in “polynomial time”: that is, there is some positive integer k such that if the size of the problem is n (in the example above, the size was the number of digits of the integer to be factorized), then the computation can be carried out in at

most n^k steps. A problem belongs to P if and only if the time taken to solve it scales up by at most a constant factor when the size of the input scales up by a constant factor. A good example of such a problem is multiplication of two n -digit numbers: if you use ordinary long multiplication, then replacing n by $2n$ increases the time taken by a factor of 4.

Suppose that you are presented with a positive integer x and told that it is a product of two primes p and q . How difficult is it to determine p and q ? Nobody knows, but one thing is easy to see: if you are told p and q , then it is not hard (for a computer, at any rate) to check that pq really does equal x . Indeed, as we have just seen, long multiplication takes polynomial time, and comparing the answer with x is even easier. The complexity class NP consists of those computational tasks for which a correct answer can be *verified* in polynomial time, even if it cannot necessarily be *found* in polynomial time. Remarkably, although this is a fundamental distinction, nobody knows how to prove that $P \neq NP$: this problem is widely considered to be the most important in theoretical computer science.

We briefly mention two other important complexity classes. PSPACE consists of all problems that can be solved using an amount of memory that grows at most polynomially with the size of the input. It turns out to be the natural class associated with reasonable computational strategies for games such as chess. The complexity class NC is the set of all Boolean functions that can be computed by a “circuit of polynomial size and depth at most a polynomial in $\log n$.” This last class is a model for the class of problems that can be solved very rapidly using parallel processing. In general, complexity classes are often surprisingly good at characterizing large families of problems with interesting and intuitively recognizable features in common. Another remarkable fact is that almost all complexity classes have “hardest problems” within them: that is, problems for which a solution can be converted into a solution for any other problem in the class. These problems are said to be *complete* for the class in question.

These issues, as well as several other complexity classes, are discussed in COMPUTATIONAL COMPLEXITY [IV.21]. A vast number of further classes can be found at

http://qwiki.stanford.edu/wiki/Complexity_Zoo#ac

along with brief definitions of each.

Continued Fractions

See THE EUCLIDEAN ALGORITHM AND
CONTINUED FRACTIONS [III.22]

III.11 Countable and Uncountable Sets

Infinite sets arise all the time in mathematics: the natural numbers, the squares, the primes, the integers, the rationals, the reals, and so on. It is often natural to try to compare the sizes of these sets: intuitively, one feels that the set of natural numbers is “smaller” than the set of integers (as it contains just the positive ones), and much larger than the set of squares (since a typical large integer is unlikely to be a square). But can we make comparisons of size in a precise way?

An obvious method of attack is to build on our intuition about finite sets. If A and B are finite sets, there are two ways we might go about comparing their sizes. One is to count their elements: we obtain two nonnegative integers m and n and just look at whether $m < n$, $m = n$, or $m > n$. But there is another important method, which does not require us to know the sizes of either A or B . This is to pair off elements from A with elements of B until one or other of the sets runs out of elements: the first one to run out is the smaller set, and if there is a dead heat, then the sets have the same size.

A suitable modification of this second method works for infinite sets as well: we can declare two sets to be of equal size if there is a one-to-one correspondence between them. This turns out to be an important and useful definition, though it does have some consequences that seem a little odd at first. For example, there is an obvious one-to-one correspondence between natural numbers and perfect squares: for each n we let n correspond to n^2 . Thus, according to this definition there are “as many” squares as there are natural numbers. Similarly, we could show that there are as many primes as natural numbers by associating n with the n th prime number.¹

What about \mathbb{Z} ? It seems that it should be “twice as large” as \mathbb{N} , but again we can find a one-to-one correspondence between them. We just list the integers in the order $0, 1, -1, 2, -2, 3, -3, \dots$ and then match the natural numbers with them in the obvious way: 0 with

1. There is a notion of “density” according to which the sets of squares and primes have density 0, the even numbers have density $\frac{1}{2}$, and so on for all sufficiently nice sets. This notion can be useful too, but it is not the notion under discussion here.

0, then 1 with 1, then 2 with -1 , 3 with 2, 4 with -2 , and so on.

An infinite set is called *countable* if it has the same size as the natural numbers. As the above example shows, this is exactly the same as saying that we can *list* the elements of the set. Indeed, if we have listed a set A as a_0, a_1, a_2, \dots , then our correspondence is just to send n to a_n . It is worth noting that there are of course many attempted listings that fail: for example, for \mathbb{Z} we might have tried $-3, -2, -1, 0, 1, 2, 3, 4, \dots$. So it is important to recognize that when we say that a set is countable we are not saying that *every* attempt to list it works, or even that the obvious attempt does: we are merely saying that there is *some* way of listing the elements. This is in complete contrast to finite sets, where if we attempt to match up two sets and find some elements of one set left over, then we know that the two sets cannot be in one-to-one correspondence. It is this difference that is mainly responsible for the “odd consequences” mentioned above.

Now that we have established that some sets that seem smaller or larger than \mathbb{N} , such as the squares or the integers, are actually countable, let us turn to a set that seems “much larger,” namely \mathbb{Q} . How could we hope to list all the rationals? After all, between any two of them you can find infinitely many others, so it seems hard not to leave some of them out when you try to list them. However, remarkable as it may seem, it *is* possible to list the rationals. The key idea is that listing the rationals whose numerator and denominator are both smaller (in modulus) than some fixed number k is easy, as there are only finitely many of them. So we go through in order: first when both numerator and denominator are at most 1, then when they are at most 2, and so on (being careful not to relist any number, so that for example $\frac{1}{2}$ should not also appear as $\frac{2}{4}$ or $\frac{3}{6}$). This leads to an ordering such as $0, 1, -1, 2, -2, \frac{1}{2}, -\frac{1}{2}, 3, -3, \frac{1}{3}, -\frac{1}{3}, 4, -4, \frac{1}{4}, -\frac{1}{4}, \frac{3}{4}, -\frac{3}{4}, \frac{4}{3}, -\frac{4}{3}, 5, -5, \dots$.

We could use the same idea to list even larger-looking sets such as, for example, the *algebraic* numbers (all real numbers, such as $\sqrt{2}$, that satisfy a polynomial equation with integer coefficients). Indeed, we note that each polynomial has only finitely many roots (which are therefore listable), so all we need to do is list the polynomials (as then we can go through them, in order, listing their roots). And we can do that by applying the same technique again: for each d we list those polynomials of degree at most d that we have not already listed, with coefficients that are at most d in modulus.

Based on the above examples, one might well guess that *every* infinite set is countable. But a beautiful argument of CANTOR [VI.54], called his “diagonal” argument, shows that the real numbers are not countable. We imagine that we have a list of all real numbers, say r_1, r_2, r_3, \dots . Our aim is to show that this list cannot possibly contain all the reals, so we wish to construct a real that is not on this list. How do we accomplish this? We have each r_i written as an infinite decimal, say, and now we define a new number s as follows. For the first digit of s (after the decimal point), we choose a digit that is not the first digit of r_1 . Note that this already guarantees that s cannot equal r_1 . (To avoid coincidences with recurring 9s and the like, it is best to choose this first digit of s not to be 0 or 9 either.) Then, for the second digit of s , we choose a digit that is not the second digit of r_2 ; this guarantees that s cannot be equal to r_2 . Continuing in this way, we end up with a real number s that is not on our list: whatever n is, the number s cannot be r_n , as s and r_n differ in the n th decimal place!

One can use similar arguments any time that we have “an infinite number of independent choices” to make in specifying an object (like the various digits of s). For example, let us use the same ideas to show that the set of all subsets of \mathbb{N} is uncountable. Suppose we have listed all the subsets as A_1, A_2, A_3, \dots . We will define a new set B that is not equal to any of the A_n . So we include the point 1 in B if and only if 1 does not belong to A_1 (this guarantees that B is not equal to A_1), and we include 2 in B if and only if 2 does not belong to A_2 , and so on. It is amusing to note that one can write this set B down as $\{n \in \mathbb{N} : n \notin A_n\}$, which shows a striking resemblance to the set in Russell’s paradox.

Countable sets are the “smallest” infinite sets. However, the set of real numbers is by no means the “largest” infinite set. Indeed, the above argument shows that no set X can be put into one-to-one correspondence with the set of all its subsets. So the set of all subsets of the real numbers is “strictly larger” than the set of real numbers, and so on.

The notion of countability is often a very fruitful one to bear in mind. For example, suppose we want to know whether or not all real numbers are algebraic. It is a genuinely hard exercise to write down a particular real that is TRANSCENDENTAL [III.43] (meaning not algebraic; see LIOUVILLE’S THEOREM AND ROTH’S THEOREM [V.25] for an idea of how it can be done), but the above notions make it utterly trivial that transcendental numbers exist. Indeed, the set of all real numbers is

uncountable but the set of algebraic numbers is countable! Furthermore, this shows that “most” real numbers are transcendental: the algebraic numbers form only a tiny proportion of the reals.

III.12 C^* -Algebras

A BANACH SPACE [III.64] is both a VECTOR SPACE [I.3 §2.3] and a METRIC SPACE [III.58], and the study of Banach spaces is therefore a mixture of linear algebra and analysis. However, one can arrive at more sophisticated mixtures of algebra and analysis if one looks at Banach spaces with more algebraic structure. In particular, while one can add two elements of a Banach space together, one cannot in general multiply them. However, sometimes one can: a vector space with a multiplicative structure is called an *algebra*, and if the vector space is also a Banach space, and if the multiplication has the property that $\|xy\| \leq \|x\| \|y\|$ for any two elements x and y , then it is called a *Banach algebra*. (This name does not really reflect historical reality, since the basic theory of Banach algebras was not worked out by Banach. A more appropriate name might have been Gelfand algebras.)

A C^* -algebra is a Banach algebra with an *involution*, which means a function that associates with each element x another element x^* in such a way that $x^{**} = x$, $\|x^*\| = \|x\|$, $(x + y)^* = x^* + y^*$, and $(xy)^* = y^*x^*$ for any two elements x and y . A basic example of a C^* -algebra is the algebra $B(H)$ of all continuous linear maps T defined on a HILBERT SPACE [III.37] H . The norm of T is defined to be the smallest constant M such that $\|Tx\| \leq M\|x\|$ for every $x \in H$, and the involution takes T to its *adjoint*. This is a map T^* that has the property that $\langle x, Ty \rangle = \langle T^*x, y \rangle$ for every x and y in H . (It can be shown that there is exactly one map with this property.) If H is finite dimensional, then T can be thought of as an $n \times n$ matrix for some n , and T^* is then the complex conjugate of the transpose of T .

A fundamental theorem of Gelfand and Naimark states that every C^* -algebra can be represented as a subalgebra of $B(H)$ for some Hilbert space H . For more information, see OPERATOR ALGEBRAS [IV.19 §3].

III.13 Curvature

If you cut an orange in half, scoop out the inside, and try to flatten one of the resulting hemispheres of peel, then you will tear it. If you try to flatten a horse’s saddle, or a soggy potato chip, then you will have the opposite

problem: this time, there is “too much” of the surface to flatten and you will have to fold it over itself. If, however, you have a roll of wallpaper and wish to flatten it, then there is no difficulty: you just unroll it. Surfaces such as spheres are said to be *positively curved*, ones with a saddle-like shape are *negatively curved*, and ones like a piece of wallpaper are *flat*.

Notice that a surface can be flat in this sense even if it does not lie in a plane. This is because curvature is defined in terms of the *intrinsic geometry* of a surface, where distance is measured in terms of paths that lie inside the surface.

There are various ways of making the above notion of curvature precise, and also quantitative, so that with each point of a surface one can associate a number that tells you “how curved” it is at that point. In order to do this, the surface must have a RIEMANNIAN METRIC [I.3 §6.10] on it, which is used to determine the lengths of paths. The notion of curvature can also be generalized to higher dimensions, so that one can talk about the curvature of a point in a d -dimensional Riemannian manifold. However, when the dimension is higher than 2, the way that the manifold can curve at a point is more complicated, and is expressed not by a single number but by the so-called *Ricci tensor*. See RICCI FLOW [III.80] for more details.

Curvature is one of the fundamental concepts of modern geometry: not only the notion just described but also various alternative definitions that measure in other ways how far a geometric object deviates from being flat. It is also an integral part of the theory of general relativity (which is discussed in GENERAL RELATIVITY AND THE EINSTEIN EQUATIONS [IV.17]).

III.14 Designs

Peter J. Cameron

Block designs were first used in the design of experiments in statistics, as a method for coping with systematic differences in the experimental material. Suppose, for example, that we want to test seven different varieties of seed in an agricultural experiment, and that we have twenty-one plots of land available for the experiment. If the plots can be regarded as identical, then the best strategy is clearly to plant three plots with each variety. Suppose, however, that the available plots are on seven farms in different regions, with three plots on each farm. If we simply plant one variety on each farm, we lose information, because we cannot distinguish systematic differences between regions from dif-

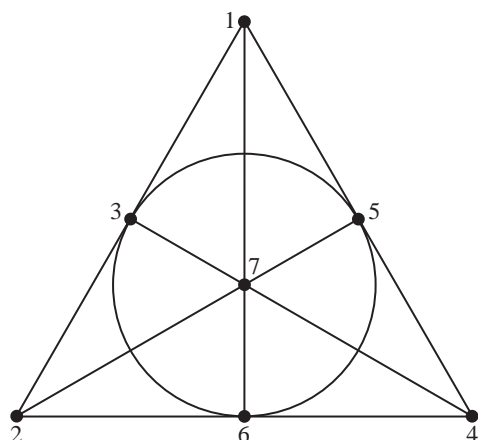


Figure 1 A block design.

ferences in the seed varieties. It is better to follow a scheme like this: plant varieties 1, 2, 3 on the first farm; 1, 4, 5 on the second; and then 1, 6, 7; 2, 4, 6; 2, 5, 7; 3, 4, 7; and 3, 5, 6. This design is represented in figure 1.

This arrangement is called a *balanced incomplete-block design*, or BIBD for short. The blocks are the sets of seed varieties used on the seven farms. The blocks are “incomplete” because not every variety can be planted on every farm; the design is “balanced” because each pair of varieties occur together in a block the same number of times (just once in this case). This is a $(7, 3, 1)$ design: there are seven varieties; each block contains three of them; and two varieties occur together in a block once. It is also an example of a finite *projective plane*. Because of the connection with geometry, varieties are usually called “points.”

Mathematicians have developed an extensive theory of BIBDs and related classes of designs. Indeed, the study of such designs predates their use in statistics. In 1847, T. P. Kirkman showed that a $(v, 3, 1)$ design exists if and only if v is congruent to 1 or 3 mod 6. (Such designs are now called *Steiner triple systems*, although Steiner did not pose the problem of their existence until 1853.)

Kirkman also posed a more difficult problem. In his own words,

Fifteen young ladies in a school walk out three abreast for seven days in succession: it is required to arrange them daily so that no two shall walk twice abreast.

The solution requires a $(15, 3, 1)$ Steiner triple system with the extra property that the thirty-five blocks can be partitioned into seven sets called “replicates,” each

replicate consisting of five blocks that partition the set of points. Kirkman himself gave a solution, but it was not until the late 1960s that Ray-Chaudhuri and Wilson showed that $(v, 3, 1)$ designs with this property exist whenever v is congruent to 3 mod 6.

For which v, k, λ do designs exist? Counting arguments show that, given k and λ , the values of v for which (v, k, λ) designs exist are restricted to certain congruence classes. (We noted above that $(v, 3, 1)$ designs exist only if v is congruent to 1 or 3 mod 6.) An asymptotic existence theory developed by Richard Wilson shows that this necessary condition is sufficient for the existence of a design, apart from finitely many exceptions, for each value of k and λ .

The concept of design has been further generalized: a t -(v, k, λ) design has the property that any t points are contained in exactly λ blocks. Luc Teirlinck showed that nontrivial t -designs exist for all t , but examples for $t > 3$ are comparatively rare.

The statisticians' concerns are a bit different. In our introductory example, if only six farms were available, we could not use a BIBD for the experiment, but would have to choose the most “efficient” possible design (allowing the most information to be obtained from the experimental results). A BIBD is most efficient if it exists; but not much is known in other cases.

There are other types of design; these can be important to statistics and also lead to new mathematics. Here, for example, is an *orthogonal array*: in any two rows of this matrix, each ordered pair of symbols from $\{0, 1, 2\}$ occurs just once:

0	0	0	1	1	1	2	2	2
0	1	2	0	1	2	0	1	2
0	1	2	1	2	0	2	0	1
0	2	1	1	0	2	2	1	0

It could be used if we had four different treatments, each of which could be applied at three different levels, and if we had nine plots available for testing.

Design theory is closely related to other combinatorial topics such as error-correcting codes; indeed, Fisher “discovered” the Hamming codes as designs five years before R. W. Hamming found them in the context of error correction. Other related subjects include packing and covering problems, and especially finite geometry, where many finite versions of classical geometries can be regarded as designs.

III.15 Determinants

The determinant of a 2×2 matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

is defined to be $ad - bc$. The determinant of a 3×3 matrix

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$$

is defined to be $aei + bfg + cdh - afh - bdi - ceg$. What do these expressions have in common, how do they generalize, and why is the generalization significant?

To begin with the first question, let us make a few simple observations. Both expressions are sums and differences of products of entries from the matrix. Each one of these products contains exactly one element from each row of the matrix and also exactly one element from each column. In both cases, a minus sign seems to attach itself to the products for which the entries selected from the matrix “slope backward” rather than forward.

Up to a point it is easy to see how to extend this definition to $n \times n$ matrices with $n \geq 4$. We simply take sums and differences of all possible products of n entries, where one entry from each row is used and one from each column. The difficulty comes in deciding which of these products to add and which to subtract. To do this we take one of the products and use it to define a permutation σ of the set $\{1, 2, \dots, n\}$ as follows. For each $i \leq n$, the product contains exactly one entry in the i th row. If it belongs to the j th column then $\sigma(i) = j$. The product is added if this permutation is even and subtracted if it is odd (see PERMUTATION GROUPS [III.70]). So, for example, the permutation corresponding to the entry afh in the 3×3 determinant above sends 1 to 1, 2 to 3, and 3 to 2. This is an odd permutation, which is why afh receives a minus sign.

We still need to explain why the particular choice of products and minus signs that we have just defined is important. The reason is that it tells us something about the effect of a matrix when it is considered as a linear map. Let A be an $n \times n$ matrix. Then, as explained in [I.3 §3.2], A specifies a linear map α from \mathbb{R}^n to \mathbb{R}^n . The determinant of A tells us what this linear map does to volumes. More precisely, if X is a subset of \mathbb{R}^n with n -dimensional volume V , then αX , the result of transforming X using the linear map α , will have volume V times the determinant of A . We could write this

symbolically as follows:

$$\text{vol}(\alpha X) = \det A \cdot \text{vol}(X).$$

For example, consider the 2×2 matrix

$$A = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}.$$

The corresponding linear map is a rotation of \mathbb{R}^2 through an angle of θ . Since rotating a shape does not affect its volume, we should expect the determinant of A to be 1, and sure enough it is $\cos^2 \theta + \sin^2 \theta$, which is 1 by Pythagoras's theorem.

The above explanation is a slight oversimplification in one respect: determinants can be negative, but clearly volumes cannot. If the determinant of a matrix is -2 , to give an example, it means that the linear map multiplies volumes by 2 but also “turns shapes inside out” by reflecting them.

Determinants have many useful properties, which become obvious once one knows the above interpretation in terms of volumes. (However, it is much less obvious that this interpretation is correct: in setting up the theory of determinants one must do some work somewhere.) Let us give three of these properties.

(i) Let V be a VECTOR SPACE [I.3 §2.3] and let $\alpha : V \rightarrow V$ be a linear map. Let $\mathbf{v}_1, \dots, \mathbf{v}_n$ be a basis of V and let A be the matrix of α with respect to this basis. Now let $\mathbf{w}_1, \dots, \mathbf{w}_n$ be another basis of V and let B be the matrix of α with respect to this different basis. Then A and B are different matrices, but since they both represent the linear map α , they must have the same effect on volumes. It follows that $\det(A) = \det(B)$. To put this another way: the determinant is better thought of as a property of linear maps rather than of matrices.

Two matrices that represent the same linear map in the above sense are called *similar*. It turns out that A and B are similar if and only if there is an invertible matrix P such that $P^{-1}AP = B$. (An $n \times n$ matrix P is *invertible* if there is a matrix Q such that PQ equals the $n \times n$ identity matrix, I_n . It turns out that QP must also equal I_n as well. If this is true, then Q is called the *inverse* of P and is denoted P^{-1} .) What we have just shown is that similar matrices have the same determinant.

(ii) If A and B are any two $n \times n$ matrices, then they represent linear maps α and β of \mathbb{R}^n . The product AB represents the linear map $\alpha\beta$: that is, the linear map that results from doing β followed by α . Since β multiplies volumes by $\det B$ and α multiplies them by $\det A$, $\alpha\beta$ multiplies them by $\det A \det B$. It follows that

$\det(AB) = \det A \det B$. (The determinant of a product equals the product of the determinants.)

(iii) If A is a linear map with determinant 0 and B is any other linear map, then AB will, by the multiplicative property just discussed, have determinant 0 as well. It follows that AB cannot equal I_n , since I_n has determinant 1. Therefore a matrix with determinant 0 is not invertible. The converse of this turns out to be true as well: a matrix with nonzero determinant is invertible. Thus, the determinant gives us a way of finding out whether a matrix can be inverted.

III.16 Differential Forms and Integration

Terence Tao

It goes without saying that integration is one of the fundamental concepts of single-variable calculus. However, there are in fact *three* concepts of integration that appear in the subject: the *indefinite integral* $\int f$ (also known as the *antiderivative*), the *unsigned definite integral* $\int_{[a,b]} f(x) dx$ (which one would use to find the area under a curve, or the mass of a one-dimensional object of varying density), and the *signed definite integral* $\int_a^b f(x) dx$ (which one would use, for instance, to compute the work required to move a particle from a to b). For simplicity we shall restrict our attention here to functions $f: \mathbb{R} \rightarrow \mathbb{R}$ that are continuous on the entire real line (and similarly, when we come to differential forms, we shall discuss only forms that are continuous on the entire domain). We shall also informally use terminology such as “infinitesimal” in order to avoid having to discuss the (routine) “epsilon-delta” analytical issues that one must resolve in order to make these integration concepts fully rigorous.

These three integration concepts are of course closely related to each other in single-variable calculus; indeed, THE FUNDAMENTAL THEOREM OF CALCULUS [I.3 §5.5] relates the signed definite integral $\int_a^b f(x) dx$ to any one of the indefinite integrals $F = \int f$ by the formula

$$\int_a^b f(x) dx = F(b) - F(a), \quad (1)$$

while the signed and unsigned integral are related by the simple identity

$$\int_a^b f(x) dx = - \int_b^a f(x) dx = \int_{[a,b]} f(x) dx, \quad (2)$$

which is valid whenever $a \leq b$.

When one moves from single-variable calculus to several-variable calculus, though, these three concepts

begin to diverge significantly from each other. The indefinite integral generalizes to the notion of a *solution to a differential equation*, or to an *integral* of a connection, VECTOR FIELD [IV.10 §5], or BUNDLE [IV.10 §5]. The unsigned definite integral generalizes to the LEBESGUE INTEGRAL [III.57], or more generally to *integration on a measure space*. Finally, the signed definite integral generalizes to the *integration of forms*, which will be our focus here. While these three concepts are still related to each other, they are not as interchangeable as they are in the single-variable setting. The integration-of-forms concept is of fundamental importance in differential topology, geometry, and physics, and also yields one of the most important examples of COHOMOLOGY [IV.10 §4], namely *de Rham cohomology*, which (roughly speaking) measures the extent to which the fundamental theorem of calculus fails in higher dimensions and on general manifolds.

To provide some motivation for the concept, let us informally revisit one of the basic applications of the signed definite integral from physics, namely computing the amount of work required to move a one-dimensional particle from point a to point b in the presence of an external field. (For example, one might be moving a charged particle in an electric field.) At the infinitesimal level, the amount of work required to move a particle from a point $x_i \in \mathbb{R}$ to a nearby point $x_{i+1} \in \mathbb{R}$ is (up to a small error) proportional to the displacement $\Delta x_i = x_{i+1} - x_i$, with the constant of proportionality $f(x_i)$ depending on the initial location x_i of the particle. Thus, the total work required for this is approximately $\sum f(x_i) \Delta x_i$. Note that we do not require x_{i+1} to be to the right of x_i , so the displacement Δx_i (or the infinitesimal work $f(x_i) \Delta x_i$) may well be negative. To return to the noninfinitesimal problem of computing the work required to move from a to b , we arbitrarily select a discrete path $x_0 = a, x_1, x_2, \dots, x_n = b$ from a to b , and approximate the work as

$$\int_a^b f(x) dx \approx \sum_{i=0}^{n-1} f(x_i) \Delta x_i. \quad (3)$$

Again, we do *not* require x_{i+1} to be to the right of x_i ; it is quite possible for the path to “backtrack” repeatedly: for instance, one might have $x_i < x_{i+1} > x_{i+2}$ for some i . However, it turns out that the effect of such backtracking eventually cancels itself out; regardless of what path we choose, the expression (3) above converges as the maximum step size tends to zero, and the

PUP: repeated cross-reference here isn't great but we think it might be the best solution. OK?

limit is the signed definite integral

$$\int_a^b f(x) dx, \quad (4)$$

provided only that the total length $\sum_{i=0}^{n-1} |\Delta x_i|$ of the path (which controls the amount of backtracking involved) stays bounded. In particular, in the case when $a = b$, so that all paths are *closed* (i.e., $x_0 = x_n$), we see that the signed definite integral is zero:

$$\int_a^a f(x) dx = 0. \quad (5)$$

From this informal definition of the signed definite integral it is obvious that we have the concatenation formula

$$\int_a^c f(x) dx = \int_a^b f(x) dx + \int_b^c f(x) dx \quad (6)$$

regardless of the relative position of the real numbers a , b , and c . In particular (setting $a = c$ and using (5)) we conclude that

$$\int_a^b f(x) dx = - \int_b^a f(x) dx.$$

Thus if we reverse a path from a to b to form a path from b to a , the sign of the integral changes. This contrasts with the *unsigned definite integral* $\int_{[a,b]} f(x) dx$, since the set $[a, b]$ of numbers between a and b is exactly the same as the set of numbers between b and a . Thus we see that paths are not quite the same as sets: they carry an *orientation* which can be reversed, whereas sets do not.

Now let us move from one-dimensional integration to higher-dimensional integration: that is, from single-variable calculus to several-variable calculus. It turns out that there are *two* objects whose dimensions may increase: the “ambient space,”¹ which will now be \mathbb{R}^n instead of \mathbb{R} , and the path, which will now become an oriented k -dimensional manifold S , over which the integration will take place. For example, if $n = 3$ and $k = 2$, then one is integrating over a surface that lives in \mathbb{R}^3 .

Let us begin with the case $n \geq 1$ and $k = 1$. Here, we will be integrating over a continuously differentiable path (or *oriented rectifiable curve*) γ in \mathbb{R}^n starting and ending at points a and b , respectively. (These points may or may not be distinct, depending on whether the path is closed or open.) From a physical point of view, we are still computing the work required to move from a to b , but now we are moving in several dimensions

instead of one. In the one-dimensional case, we did not need to specify exactly which path we used to get from a to b , because all backtracking canceled itself out. However, in higher dimensions, the exact choice of the path γ becomes important.

Formally, a path from a to b can be described (or *parametrized*) as a continuously differentiable function γ from the unit interval $[0, 1]$ to \mathbb{R}^n such that $\gamma(0) = a$ and $\gamma(1) = b$. For instance, the line segment from a to b can be parametrized as $\gamma(t) = (1 - t)a + tb$. This segment also has many other parametrizations, such as $\tilde{\gamma}(t) = (1 - t^2)a + t^2b$; however, as in the one-dimensional case, the exact choice of parametrization does not ultimately influence the integral. On the other hand, the reverse line segment $(-\gamma)(t) = ta + (1 - t)b$ from b to a is a genuinely different path; the integral along $-\gamma$ will turn out to be the negative of the integral along γ .

As in the one-dimensional case, we will need to approximate the continuous path γ by a discrete path $x_0 = \gamma(t_0)$, $x_1 = \gamma(t_1)$, $x_2 = \gamma(t_2)$, ..., $x_n = \gamma(t_n)$, where $\gamma(t_0) = a$ and $\gamma(t_1) = b$. Again, we allow some backtracking: t_{i+1} is not necessarily larger than t_i . The displacement $\Delta x_i = x_{i+1} - x_i \in \mathbb{R}^n$ from x_i to x_{i+1} is now a *vector* rather than a scalar. (Indeed, with an eye on the generalization to manifolds, one should think of Δx_i as an infinitesimal *tangent vector* to the ambient space \mathbb{R}^n at the point x_i .) In the one-dimensional case, we converted the scalar displacement Δx_i into a new number $f(x_i)\Delta x_i$, which was linearly related to the original displacement by a proportionality constant $f(x_i)$ that depended on the position x_i . In higher dimensions, we again have a linear dependence, but this time, since the displacement is a vector, we must replace the simple constant of proportionality by a *linear transformation* ω_{x_i} from \mathbb{R}^n to \mathbb{R} . Thus, $\omega_{x_i}(\Delta x_i)$ represents the infinitesimal “work” required to move from x_i to x_{i+1} . In technical terms, ω_{x_i} is a *linear functional* on the space of tangent vectors at x_i , and is thus a *cotangent vector* at x_i . By analogy with (3), the net work $\int_\gamma \omega$ required to move from a to b along the path γ is approximated by

$$\int_\gamma \omega \approx \sum_{i=0}^{n-1} \omega_{x_i}(\Delta x_i). \quad (7)$$

As in the one-dimensional case, one can show that the right-hand side of (7) converges if the maximum step size $\sup_{0 \leq i \leq n-1} |\Delta x_i|$ of the path converges to zero and the total length $\sum_{i=0}^{n-1} |\Delta x_i|$ of the path stays

1. We will start with integration on Euclidean spaces \mathbb{R}^n for simplicity, although the true power of the integration-of-forms concept is much more apparent when we integrate on more general spaces, such as abstract n -dimensional manifolds.

bounded. The limit is written as $\int_Y \omega$. (Recall that we are restricting our attention to continuous functions. The existence of this limit uses the continuity of ω .)

The object ω , which continuously assigns² a cotangent vector to each point in \mathbb{R}^n , is called a *1-form*, and (7) leads to a recipe for integrating any 1-form ω on a path γ . That is, to shift the emphasis slightly, it allows us to integrate the path γ “against” the 1-form ω . Indeed, it is useful to think of this integration as a *binary* operation (similar in some ways to the dot product) which takes the curve γ and the form ω as inputs, and returns a scalar $\int_Y \omega$ as output. There is in fact a “duality” between curves and forms; compare, for instance, the identity

$$\int_Y (\omega_1 + \omega_2) = \int_Y \omega_1 + \int_Y \omega_2,$$

which expresses (part of) the fundamental fact that integration of forms is a linear operation, with the identity

$$\int_{\gamma_1 + \gamma_2} \omega = \int_{\gamma_1} \omega + \int_{\gamma_2} \omega,$$

which generalizes (6) whenever the initial point of γ_2 is the final point of γ_1 , where $\gamma_1 + \gamma_2$ is the *concatenation* of γ_1 and γ_2 .³

Recall that if f is a differentiable function from \mathbb{R}^n to \mathbb{R} , then its derivative at a point x is a linear map from \mathbb{R}^n to \mathbb{R} (see [I.3 §5.3]). If f is continuously differentiable, then this linear map depends continuously on x , and can therefore be thought of as a 1-form, which we denote by df , writing df_x for the derivative at x . This 1-form can be characterized as the unique 1-form such that one has the approximation

$$f(x + v) \approx f(v) + df_x(v)$$

for all infinitesimal v . (More rigorously, the condition is that $|f(x + v) - f(v) - df_x(v)|/|v| \rightarrow 0$ as $v \rightarrow 0$.)

The fundamental theorem of calculus (1) now generalizes to

$$\int_Y df = f(b) - f(a) \quad (8)$$

whenever γ is any oriented curve from a point a to a point b . In particular, if γ is closed, then $\int_Y df = 0$. Note that in order to interpret the left-hand side of the above equation, we are regarding it as a particular example of

an integral of the form $\int_Y \omega$: in this case, ω happens to be the form df . Note also that, with this interpretation, df has an independent meaning (it is a 1-form) even if it does not appear under an integral sign.

A 1-form whose integral against every sufficiently small⁴ closed curve vanishes is called *closed*, while a 1-form that can be written as df for some continuously differentiable function is called *exact*. Thus, the fundamental theorem implies that every exact form is closed. This turns out to be a general fact, valid for all manifolds. Is the converse true: that is, is every closed form exact? If the domain is a Euclidean space, or indeed any other *simply connected* manifold, then the answer is yes (this is a special case of the *Poincaré lemma*), but it is not true for general domains. In modern terminology, this demonstrates that the de Rham cohomology of such domains can be nontrivial.

As we have just seen, a 1-form can be thought of as an object ω that associates with each path γ a scalar, which we denote by $\int_Y \omega$. Of course, ω is not just any old function from paths to scalars: it must satisfy the concatenation and reversing rules discussed earlier, and this, together with our continuity assumptions, more or less forces it to be associated with some kind of continuously varying linear function that can be used, in combination with γ , to define an integral. Now let us see if we can generalize this basic idea from paths to integration on k -dimensional sets with $k > 1$. For simplicity we shall stick to the two-dimensional case, that is, to integration of forms on (oriented) surfaces in \mathbb{R}^n , since this already illustrates many features of the general case.

Physically, such integrals arise when one is computing a *flux* of some field (e.g., a magnetic field) across a surface. We parametrized one-dimensional oriented curves as continuously differentiable functions γ from the interval $[0, 1]$ to \mathbb{R}^n . It is thus natural to parametrize two-dimensional oriented surfaces as continuously differentiable functions ϕ defined on the unit square $[0, 1]^2$. This does not in fact cover all possible surfaces one wishes to integrate over, but it turns out that one can cut up more general surfaces into pieces that can be parametrized using “nice” domains such as $[0, 1]^2$.

In the one-dimensional case, we cut up the oriented interval $[0, 1]$ into infinitesimal oriented intervals from t_i to $t_{i+1} = t_i + \Delta t$, which led to infinitesimal curves

2. More precisely, one can think of ω as a *section* of the *cotangent bundle*.

3. This duality is best understood using the abstract, and much more general, formalism of homology and cohomology. In particular, one can remove the requirement that γ_2 begins where γ_1 leaves off by generalizing the notion of an integral to cover not just integration on paths, but also integration on *formal sums or differences* of paths. This makes the duality between curves and forms more symmetric.

4. The precise condition needed is that the curve should be *contractible*, which means that it can be continuously shrunk down to a point.

from $x_i = y(t_i)$ to $x_{i+1} = y(t_{i+1}) = x_i + \Delta x_i$. Note that Δx_i and Δt are related by the approximation $\Delta x_i \approx y'(t_i)\Delta t_i$. In the two-dimensional case, we will cut up the unit square $[0, 1]^2$ into infinitesimal squares in an obvious way.⁵ A typical one of these will have corners of the form (t_1, t_2) , $(t_1 + \Delta t, t_2)$, $(t_1, t_2 + \Delta t)$, $(t_1 + \Delta t, t_2 + \Delta t)$. The surface described by ϕ can then be partitioned into regions with corners $\phi(t_1, t_2)$, $\phi(t_1 + \Delta t, t_2)$, $\phi(t_1, t_2 + \Delta t)$, $\phi(t_1 + \Delta t, t_2 + \Delta t)$, each of which carries an orientation. Since ϕ is differentiable, it is approximately linear at small distance scales, so this region is approximately an oriented parallelogram in \mathbb{R}^n with corners x , $x + \Delta_1 x$, $x + \Delta_2 x$, $x + \Delta_1 x + \Delta_2 x$, where $x = (t_1, t_2)$ and $\Delta_1 x$ and $\Delta_2 x$ are the infinitesimal vectors

$$\Delta_1 x = \frac{\partial \phi}{\partial t_1}(t_1, t_2)\Delta t, \quad \Delta_2 x = \frac{\partial \phi}{\partial t_2}(t_1, t_2)\Delta t.$$

Let us refer to this object as the infinitesimal parallelogram with *dimensions* $\Delta_1 x \wedge \Delta_2 x$ and *base point* x . For now, we will think of the symbol “ \wedge ” as a mere notational convenience and not try to interpret it. In order to integrate in a manner analogous with integration on curves, we now need some sort of functional ω_x at this base point that depends continuously on x . This functional should take the above infinitesimal parallelogram and return an infinitesimal number $\omega_x(\Delta_1 x \wedge \Delta_2 x)$, which one can think of as the amount of “flux” passing through this parallelogram.

As in the one-dimensional case, we expect ω_x to have certain properties. For instance, if you double $\Delta_1 x$, you double one of the sides of the infinitesimal parallelogram, so (by the continuity of ω) the “flux” passing through the parallelogram should double. More generally, $\omega_x(\Delta_1 x \wedge \Delta_2 x)$ should depend linearly on each of $\Delta_1 x$ and $\Delta_2 x$: in other words, it is *bilinear*. (This generalizes the linear dependence in the one-dimensional case.)

Another important property is that

$$\omega_x(\Delta_2 x \wedge \Delta_1 x) = -\omega_x(\Delta_1 x \wedge \Delta_2 x). \quad (9)$$

That is, the bilinear form ω_x is *antisymmetric*. Again, this has an intuitive explanation: the parallelogram represented by $\Delta_2 x \wedge \Delta_1 x$ is the same as that represented by $\Delta_1 x \wedge \Delta_2 x$ except that it has had its orientation reversed, so the “flux” now counts negatively where it used to count positively, and vice versa. Another way

of seeing this is to note that if $\Delta_1 x = \Delta_2 x$, then the parallelogram is degenerate and there should be no flux. Antisymmetry follows from this and the bilinearity. A 2-form ω is a continuous assignment of a functional ω_x with these properties to each point x .

If ω is a 2-form and $\phi : [0, 1]^2 \rightarrow \mathbb{R}^n$ is a continuously differentiable function, we can now define the integral $\int_\phi \omega$ of ω “against” ϕ (or, more precisely, the integral against the image under ϕ of the oriented square $[0, 1]^2$) by the approximation

$$\int_\phi \omega \approx \sum_i \omega_{x_i}(\Delta x_{1,i} \wedge \Delta x_{2,i}), \quad (10)$$

where the image of ϕ is (approximately) partitioned into parallelograms of dimensions $\Delta x_{1,i} \wedge \Delta x_{2,i}$ based at points x_i . We do not need to decide what order these parallelograms should be arranged in, because addition is both commutative and associative. One can show that the right-hand side of (10) converges to a unique limit as one makes the partition of parallelograms “increasingly fine,” though we will not make this precise here.

We have thus shown how to integrate 2-forms against oriented two-dimensional surfaces. More generally, one can define the concept of a k -form on an n -dimensional manifold (such as \mathbb{R}^n) for any $0 \leq k \leq n$ and integrate this against an oriented k -dimensional surface in that manifold. For instance, a 0-form on a manifold X is the same thing as a scalar function $f : X \rightarrow \mathbb{R}$, whose integral on a positively oriented point x (which is zero dimensional) is $f(x)$, and on a negatively oriented point x is $-f(x)$. A k -form tells us how to assign a value to an infinitesimal k -dimensional parallelepiped with dimensions $\Delta x_1 \wedge \cdots \wedge \Delta x_k$, and hence to a portion of k -dimensional “surface,” in much the same way as we have seen when $k = 2$. By convention, if $k \neq k'$, the integral of a k -dimensional form on a k' -dimensional surface is understood to be zero. We refer to 0-forms, 1-forms, 2-forms, etc. (and formal sums and differences thereof), collectively as *differential forms*.

There are three fundamental operations that one can perform on scalar functions: addition $(f, g) \mapsto f + g$, pointwise product $(f, g) \mapsto fg$, and differentiation $f \mapsto df$, although the last of these is not especially useful unless f is continuously differentiable. These operations have various relationships with each other. For instance, the product is *distributive* over addition,

$$f(g + h) = fg + fh,$$

5. One could also use infinitesimal oriented rectangles, parallelograms, triangles, etc.; this leads to an equivalent concept of the integral.

and differentiation is a *derivation* with respect to the product:

$$d(fg) = (df)g + f(dg).$$

It turns out that one can generalize all three of these operations to differential forms. Adding a pair of forms is easy: if ω and η are two k -forms and $\phi : [0, 1]^k \rightarrow \mathbb{R}^n$ is a continuously differentiable function, then $\int_\phi(\omega + \eta)$ is defined to be $\int_\phi \omega + \int_\phi \eta$. One multiplies forms using the so-called *wedge product*. If ω is a k -form and η is an l -form, then $\omega \wedge \eta$ is a $(k + l)$ -form. Roughly speaking, given a $(k + l)$ -dimensional infinitesimal parallelepiped with base point x and dimensions $\Delta x_1 \wedge \cdots \wedge \Delta x_{k+l}$, one evaluates ω and η at the parallelepipeds with base point x and dimensions $\Delta x_1 \wedge \cdots \wedge \Delta x_k$ and $\Delta x_{k+1} \wedge \cdots \wedge \Delta x_{k+l}$, respectively, and multiplies the results together.

As for differentiation, if ω is a continuously differentiable k -form, then its derivative $d\omega$ is a $k + 1$ -form that measures something like the “rate of change” of ω . To see what this might mean, and in particular to see why $d\omega$ is a $k + 1$ form, let us think how we might answer a question of the following kind. We are given a spherical surface in \mathbb{R}^3 and a flow, and we would like to know the net flux out of the surface: that is, the difference between the amount of flux coming out and the amount going in. One way to do this would be to approximate the surface of the sphere by a union of tiny parallelograms, to measure the flux through each one, and to take the sum of all these fluxes. Another would be to approximate the solid sphere by a union of tiny parallelepipeds, to measure the *net* flux out of each of these, and to add up the results. If a parallelepiped is small enough, then we can closely approximate the net flux out of it by looking at the difference, for each pair of opposite faces, between the amount coming out of the parallelepiped through one and the amount going into it through the other, and this will depend on the rate of change of the 2-form.

The process of summing up the net fluxes out of the parallelepipeds is more rigorously described as integrating a 3-form over the solid sphere. In this way, one can see that it is natural to expect that information about how a 2-form varies should be encapsulated in a 3-form.

The exact construction of these operations requires a little bit of algebra and is omitted here. However, we remark that they obey similar laws to their scalar counterparts, except that there are some sign changes that are ultimately due to the antisymmetry (9). For

instance, if ω is a k -form and η is an l -form, the commutative law for multiplication becomes

$$\omega \wedge \eta = (-1)^{kl} \eta \wedge \omega,$$

basically because kl swaps are needed to interchange k dimensions with l dimensions; and the derivation rule for differentiation becomes

$$d(\omega \wedge \eta) = (d\omega) \wedge \eta + (-1)^k \omega \wedge (d\eta).$$

Another rule is that the differentiation operator d is nilpotent:

$$d(d\omega) = 0. \quad (11)$$

This may seem rather unintuitive, but it is fundamentally important. To see why it might be expected, let us think about differentiating a 1-form twice. The original 1-form associates a scalar with each small line segment. Its derivative is a 2-form that associates a scalar with each small parallelogram. This scalar essentially measures the sum of the scalars given by the 1-form as you go around the four edges of the parallelogram, though to get a sensible answer when you pass to the limit you have to divide by the area of the parallelogram. If we now repeat the process, we are looking at a sum of the six scalars associated with the six faces of a parallelepiped. But each of these scalars in turn comes from a sum of the scalars associated with the four directed edges around the corresponding face, and each edge is therefore counted twice (as it belongs to two faces), once in each direction. Therefore, the contributions from each edge cancel and the sum of all contributions is zero.

The description given earlier of the relationship between integrating a 2-form over the surface of a sphere and integrating its derivative over the solid sphere can be thought of as a generalization of the fundamental theorem of calculus, and can itself be generalized considerably: *Stokes's theorem* is the assertion that

$$\int_S d\omega = \int_{\partial S} \omega \quad (12)$$

for any oriented manifold S and form ω , where ∂S is the oriented boundary of S (which we will not define here). Indeed one can view this theorem as a definition of the derivative operation $\omega \mapsto d\omega$; thus, differentiation is the *adjoint* of the boundary operation. (For instance, the identity (11) is dual to the geometric observation that the boundary ∂S of an oriented manifold itself has no boundary: $\partial(\partial S) = \emptyset$.) As a particular case of Stokes's theorem, we see that $\int_S d\omega = 0$ whenever S is a *closed* manifold, i.e., one with no boundary. This observation

lets one extend the notions of closed and exact forms to general differential forms, which (together with (11)) allows one to fully set up *de Rham cohomology*.

We have already seen that 0-forms can be identified with scalar functions. Also, in Euclidean spaces one can use the inner product to identify linear functionals with vectors, and therefore 1-forms can be identified with vector fields. In the special (but very physical) case of three-dimensional Euclidean space \mathbb{R}^3 , 2-forms can *also* be identified with vector fields via the famous *right-hand rule*,⁶ and 3-forms can be identified with scalar functions by a variant of this rule. (This is an example of a concept known as *Hodge duality*.) In this case, the differentiation operation $\omega \mapsto d\omega$ can be identified with the *gradient* operation $f \mapsto \nabla f$ when ω is a 0-form, with the *curl* operation $X \mapsto \nabla \times X$ when ω is a 1-form, and with the *divergence* operation $X \mapsto \nabla \cdot X$ when ω is a 2-form. Thus, for instance, the rule (11) implies that $\nabla \times \nabla f = 0$ and $\nabla \cdot (\nabla \times X) = 0$ for any suitably smooth scalar function f and vector field X , while various cases of Stokes's theorem (12), with this interpretation, become the various theorems about integrals of curves and surfaces in three dimensions that you may have seen referred to as "the divergence theorem," "Green's theorem," and "Stokes's theorem" in a course on several-variable calculus.

Just as the signed definite integral is connected to the unsigned definite integral in one dimension via (2), there is a connection between integration of differential forms and the Lebesgue (or Riemann) integral. On the Euclidean space \mathbb{R}^n one has the n standard coordinate functions $x_1, x_2, \dots, x_n : \mathbb{R}^n \rightarrow \mathbb{R}$. Their derivatives dx_1, \dots, dx_n are then 1-forms on \mathbb{R}^n . Taking their wedge product, one obtains an n -form $dx_1 \wedge \dots \wedge dx_n$. We can multiply this with any (continuous) scalar function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ to obtain another n -form $dx_1 \wedge \dots \wedge dx_n$. If Ω is any open bounded domain in \mathbb{R}^n , we then have the identity

$$\int_{\Omega} f(x) dx_1 \wedge \dots \wedge dx_n = \int_{\Omega} f(x) dx,$$

where on the left-hand side we have an integral of a differential form (with Ω viewed as a positively oriented n -dimensional manifold) and on the right-hand side we have the Riemann or Lebesgue integral of f on Ω . If we give Ω the negative orientation, we have to reverse

the sign of the left-hand side. This correspondence generalizes (2).

There is one last operation on forms that is worth pointing out. Suppose we have a continuously differentiable map $\Phi : X \rightarrow Y$ from one manifold to another (we allow X and Y to have different dimensions). Then of course every point x in X *pushes forward* to a point $\Phi(x)$ in Y . Similarly, if we let $v \in T_x X$ be an infinitesimal tangent vector to X based at x , then this tangent vector also *pushes forward* to a tangent vector $\Phi_* v \in T_{\Phi(x)}(Y)$ based at $\Phi(x)$; informally speaking, $\Phi_* v$ can be defined by requiring the infinitesimal approximation $\Phi(x + v) = \Phi(x) + \Phi_* v$. One can write $\Phi_* v = D\Phi(x)(v)$, where $D\Phi : T_x X \rightarrow T_{\Phi(x)} Y$ is the *derivative* of the several-variable map Φ at x . Finally, any k -dimensional oriented manifold S in X also pushes forward to a k -dimensional oriented manifold $\Phi(S)$ in Y , although in some cases (e.g., if the image of Φ has dimension less than k) this pushed-forward manifold may be degenerate.

We have seen that integration is a duality pairing between manifolds and forms. Since manifolds push forward under Φ from X to Y , we expect forms to *pull back* from Y to X . Indeed, given any k -form ω on Y , we can define the *pull-back* $\Phi^* \omega$ as the unique k -form on X such that we have the *change-of-variables formula*

$$\int_{\Phi(S)} \omega = \int_S \Phi^* (\omega).$$

In the case of 0-forms (i.e., scalar functions), the pull-back $\Phi^* f : X \rightarrow \mathbb{R}$ of a scalar function $f : Y \rightarrow \mathbb{R}$ is given explicitly by $\Phi^* f(x) = f(\Phi(x))$, while the pull-back of a 1-form ω is given explicitly by the formula

$$(\Phi^* \omega)_x(v) = \omega_{\Phi(x)}(\Phi_* v).$$

Similar definitions can be given for other differential forms. The pull-back operation enjoys several nice properties: for instance, it respects the wedge product,

$$\Phi^* (\omega \wedge \eta) = (\Phi^* \omega) \wedge (\Phi^* \eta),$$

and the derivative,

$$d(\Phi^* \omega) = \Phi^* (d\omega).$$

By using these properties, one can recover rather painlessly the change-of-variables formulas in several-variable calculus. Moreover, the whole theory carries effortlessly over from Euclidean spaces to other manifolds. It is because of this that the theory of differential forms and integration is an indispensable tool in the modern study of manifolds, and especially in DIFFERENTIAL TOPOLOGY [IV.9].

6. This is an entirely arbitrary convention; one could just as easily have used the left-hand rule to provide this identification, and apart from some harmless sign changes here and there, one gets essentially the same theory as a consequence.

III.17 Dimension

What is the difference between a two-dimensional set and a three-dimensional set? A rough answer that one might give is that a two-dimensional set lives inside a plane, while a three-dimensional set fills up a portion of space. Is this a good answer? For many sets it does seem to be: triangles, squares, and circles can be drawn in a plane, while tetrahedra, cubes, and spheres cannot. But how about the surface of a sphere? This we would normally think of as two dimensional, contrasting it with the solid sphere, which is three dimensional. But the surface of a sphere does not live inside a plane.

Does this mean that our rough definition was incorrect? Not exactly. From the perspective of linear algebra, the set $\{(x, y, z) : x^2 + y^2 + z^2 = 1\}$, which is the surface of a sphere of radius 1 in \mathbb{R}^3 centered at the origin, is three dimensional, precisely because it is not contained in a plane. (One can express this in algebraic language by saying that the affine subspace generated by the sphere is the whole of \mathbb{R}^3 .) However, this sense of “three dimensional” does not do justice to the rough idea that the surface of a sphere has no thickness. Surely there ought to be another sense of dimension in which the surface of a sphere is two dimensional?

As this example illustrates, dimension, though very important throughout mathematics, is not a single concept. There turn out to be many natural ways of generalizing our ideas about the dimensions of simple sets such as squares and cubes, and they are often incompatible with one another, in the sense that the dimension of a set may vary according to which definition you use. The remainder of this article will set out a few different definitions.

One very basic idea we have about the dimension of a set is that it is “the number of coordinates you need to specify a point.” We can use this to justify our instinct that the surface of a sphere is two dimensional: you can specify any point by giving its longitude and latitude. It is a little tricky to turn this idea into a rigorous mathematical definition because you can in fact specify a point of the sphere by means of just *one* number if you do not mind doing it in a highly artificial way. This is because you can take any two numbers and interleave the digits to form a single number from which the original two numbers can be recovered. For instance, from the two numbers $\pi = 3.141592653\dots$ and $e = 2.718281828\dots$ you can form the number $32.174118529821685238\dots$, and by taking alternate

digits you get back π and e again. It is even possible to find a *continuous* function f from the closed interval $[0, 1]$ (that is, the set of all real numbers between 0 and 1, inclusive) to the surface of a sphere that takes every value.

We therefore have to decide what we mean by a “natural” coordinate system. One way of making this decision leads to the definition of a *manifold*, a very important concept that is discussed in [I.3 §6.9] and also in DIFFERENTIAL TOPOLOGY [IV.9]. This is based on the idea that every point in the sphere is contained in a neighborhood N that “looks like” a piece of the plane, in the sense that there is a “nice” one-to-one correspondence ϕ between N and a subset of the Euclidean plane \mathbb{R}^2 . Here, “nice” can have different meanings: typical ones are that ϕ and its inverse should both be continuous, or differentiable, or infinitely differentiable.

Thus, the intuitive notion that a d -dimensional set is one where you need d numbers to specify a point can be developed into a rigorous definition that tells us, as we had hoped, that the surface of a sphere is two dimensional. Now let us take another intuitive notion and see what we can get from it.

Suppose I want to cut a piece of paper into two pieces. The boundary that separates the pieces will be a curve, which we would normally like to think of as one dimensional. Why is it one dimensional? Well, we could use the same reasoning: if you cut a curve into two pieces then the part where the two pieces meet each other is a single point (or pair of points if the curve is a loop), which is zero dimensional. That is, there appears to be a sense in which a $(d - 1)$ -dimensional set is needed if you want to cut a d -dimensional set into two.

Let us try to be slightly more precise about this idea. Suppose that X is a set and x and y are points in X . Let us call a set Y a *barrier* between x and y if there is no continuous path from x to y that avoids Y . For example, if X is a solid sphere of radius 2, x is the center of X , and y is a point on the boundary of X , then one possible barrier between x and y is the surface of a sphere of radius 1. With this terminology in place, we can make the following inductive definition. A finite set is zero dimensional, and in general we say that X is *at most d dimensional* if between any two points in X there is a barrier that is at most $(d - 1)$ dimensional. We also say that X is *d dimensional* if it is at most d dimensional but not at most $(d - 1)$ dimensional.

The above definition makes sense, but it runs into difficulties: one can construct a pathological set X that acts as a barrier between any two points in the plane,

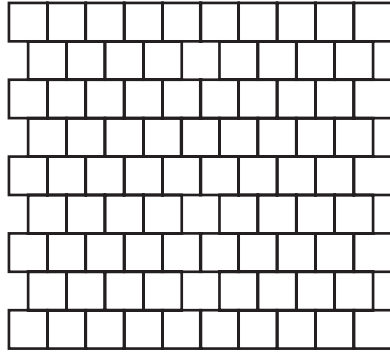


Figure 1 How to cover with squares so that no four overlap.

but contains no segment of any curve. This makes X zero dimensional and therefore makes the plane one dimensional, which is not satisfactory. A small modification to the above definition eliminates such pathologies and gives a definition that was put forward by BROUWER [VI.75]. A complete METRIC SPACE [III.58] X is said to have dimension at most d if, given any pair of disjoint closed sets A and B , you can find disjoint open sets U and V with $A \subset U$ and $B \subset V$ such that the complement Y of $U \cup V$ (that is, everything in X that does not belong to either U or V) has dimension at most $d - 1$. The set Y is the barrier—the main difference is that we have now asked for it to be closed. The induction starts with the empty set, which has dimension -1 . Brouwer's definition is known as the *inductive dimension* of a set.

Here is another basic idea that leads to a useful definition of dimension, proposed by LEBESGUE [VI.72]. Suppose you want to cover an open interval of real numbers (that is, an interval that does not contain its endpoints) with shorter open intervals. Then you will be forced to make the shorter ones overlap, but you can do it in such a way that no point is contained in more than two of your intervals: just start each new interval close to the end of the previous one.

Now suppose that you want to cover an open square (that is, one that does not contain its boundary) with smaller open squares. Again you will be forced to make the smaller squares overlap, but this time the situation is slightly worse: some points will have to be contained in three squares. However, if you take squares arranged like bricks, as in figure 1, and expand them slightly, then you can do the covering in such a way that no four squares overlap. In general, it seems that to cover a typical d -dimensional set with small open sets, you

need to have overlaps of $d + 1$ sets but you do not need to have overlaps greater than this.

The precise definition that this leads to is surprisingly general: it makes sense not just for subsets of \mathbb{R}^n but even for an arbitrary TOPOLOGICAL SPACE [III.92]. We say that a set X is *at most d dimensional* if, however you cover X with a finite collection of open sets U_1, \dots, U_n , you can find a finite collection of open sets V_1, \dots, V_m with the following properties:

- (i) the sets V_i also cover the whole of X ;
- (ii) every V_i is a subset of at least one U_i ;
- (iii) no point is contained in more than $d + 1$ of the V_i .

If X is a metric space, then we can choose our U_i to have small diameter, thereby forcing the V_i to be small. So this definition is basically saying that it is possible to cover X with open sets with no $d + 2$ of them overlapping, and that these open sets can be as small as you like.

As with inductive dimension, we then define the *dimension* of X to be the smallest d such that X is at most d dimensional. And again it can be shown that this definition assigns the “correct” dimension to the familiar shapes of elementary geometry.

A fourth intuitive idea leads to concepts known as *homological* and *cohomological* dimension. Associated with any suitable topological space X , such as a manifold, are sequences of groups known as HOMOLOGY AND COHOMOLOGY GROUPS [IV.10§4]. Here we will discuss homology groups, but a very similar discussion is possible for cohomology. Roughly speaking, the n th homology group tells you how many interestingly different continuous maps there are from closed n -dimensional manifolds M to X . If X is a manifold of dimension less than n , then it can be shown that the n th homology group is trivial: in a sense, there is not enough room in X to define any map that is interestingly different from a constant map. On the other hand, the n th homology group of the n -sphere itself is \mathbb{Z} , which says that one can classify the maps from the n -sphere to itself by means of an integer parameter.

It is therefore tempting to say that a space is at least n dimensional if there is room inside it for interesting maps from n -dimensional manifolds. This thought leads to a whole class of definitions. The homological dimension of a structure X is defined to be the largest n for which some substructure of X has a nontrivial n th homology group. (It is necessary to consider substructures, because homology groups can also be trivial

when there is too *much* room: it then becomes easy to deform a continuous map and show that it is equivalent to a constant map.) However, homology is a very general concept and there are many different homology theories, so there are many different notions of homological dimension. Some of these are geometric, but there are also homology theories for algebraic structures: for example, using suitable theories, one can define the homological dimension of algebraic structures such as RINGS [III.83 §1] or GROUPS [I.3 §2.1]. This is a very good example of geometrical ideas having an algebraic payoff.

Now let us turn to a fifth and final (for this article at least) intuitive idea about dimension, namely the way it affects how we measure *size*. If you want to convey how big a shape X is, then a good way of doing so is to give the length of X if X is one dimensional, the area if it is two dimensional, and the volume if it is three dimensional. Of course, this presupposes that you already know what the dimension is, but, as we shall see, there is a way of deciding which measure is the most appropriate *without* determining the dimension in advance. Then the tables are turned: we can actually *define* the dimension to be the number that corresponds to the best measure.

To do this, we use the fact that length, area, and volume scale in different ways when you expand a shape. If you take a curve and expand it by a factor of 2 (in all directions), then its length doubles. More generally, if you expand by a factor of C , then the length multiplies by C . However, if you take a two-dimensional shape and expand it by C , then its area multiplies by C^2 . (Roughly speaking, this is because each little portion of the shape expands by C “in two directions” so you have to multiply the area by C twice.) And the volume of a three-dimensional shape multiplies by C^3 : for instance, the volume of a sphere of radius 3 is twenty-seven times the volume of a sphere of radius 1.

It may look as though we still have to decide in advance whether we will talk about length, area, or volume before we can even begin to think about how the measurement scales when we expand the shape. But this is not the case. For instance, if we expand a square by a factor of 2, then we obtain a new square that can be divided up into four congruent copies of the original square. So, without having decided in advance that we are talking about area, we can say that the size of the new square is four times that of the old square.

This observation has a remarkable consequence: there are sets to which it is natural to assign a dimen-

sion that is not an integer! Perhaps the simplest example is a famous set first defined by CANTOR [VI.54] and now known as the *Cantor set*. This set is produced as follows. You start with the closed interval $[0, 1]$, and call it X_0 . Then you form a set X_1 by removing the middle third of X_0 : that is, you remove all points between $\frac{1}{3}$ and $\frac{2}{3}$, but leave $\frac{1}{3}$ and $\frac{2}{3}$ themselves. So X_1 is the union of the closed intervals $[0, \frac{1}{3}]$ and $[\frac{2}{3}, 1]$. Next, you remove the middle thirds of these two closed intervals to produce a set X_2 , so X_2 is the union of the intervals $[0, \frac{1}{9}]$, $[\frac{2}{9}, \frac{1}{3}]$, $[\frac{2}{3}, \frac{7}{9}]$, and $[\frac{8}{9}, 1]$.

In general, X_n is a union of closed intervals, and X_{n+1} is what you get by removing the middle thirds of each of these intervals—so X_{n+1} consists of twice as many intervals as X_n , but they are a third of the size. Once you have produced the sequence X_0, X_1, X_2, \dots , you define the Cantor set to be the intersection of all the X_i : that is, all the real numbers that remain, no matter how far you go with the process of removing middle thirds of intervals. It is not hard to show that these are precisely the numbers whose ternary expansions consist just of 0s and 2s. (There are some numbers that have two different ternary expansions. For instance, $\frac{1}{3}$ can be written either as 0.1 or as 0.22222... In such cases we take the recurring expansion rather than the terminating one. So $\frac{1}{3}$ belongs to the Cantor set.) Indeed, when you remove middle thirds for the n th time, you are removing all numbers that have a 1 in the n th place after the “decimal” (in fact, ternary) point.

The Cantor set has many interesting properties. For example, it is UNCOUNTABLE [III.11], but it also has MEASURE [III.57] zero. Briefly, the first of these assertions follows from the fact that there is a different element of the Cantor set for every subset A of the natural numbers (just take the ternary number $0.a_1a_2a_3\dots$, where $a_i = 2$ whenever $i \in A$ and $a_i = 0$ otherwise), and there are uncountably many subsets of the natural numbers. To justify the second, note that the total length of the intervals making up X_n is $(\frac{2}{3})^n$ (since one removes a third of X_{n-1} to produce X_n). Since the Cantor set is contained in every X_n , its measure must be smaller than $(\frac{2}{3})^n$, whatever n is, which means that it must be zero. Thus, the Cantor set is very large in one respect and very small in another.

A further property of the Cantor set is that it is *self-similar*. The set X_1 consists of two intervals, and if you look at just one of these intervals as the middle thirds are repeatedly removed, then what you see is just like the construction of the whole Cantor set, but scaled down by a factor of 3. That is, the Cantor set consists

of two copies of itself, each scaled down by a factor of 3. From this we deduce the following statement: if you expand the Cantor set by a factor of 3, then you can divide the expanded set up into two congruent copies of the original, so it is “twice as big.”

What consequence should this have for the dimension of the Cantor set? Well, if the dimension is d , then the expanded set ought to be 3^d times as big. Therefore, 3^d should equal 2. This means that d should be $\log 2 / \log 3$, which is roughly 0.63.

Once one knows this, the mystery of the Cantor set is lessened. As we shall see in a moment, a theory of fractional dimension can be developed with the useful property that a countable union of sets of dimension at most d has dimension at most d . Therefore, the fact that the Cantor set has dimension greater than 0 implies that it cannot be countable (since single points have dimension 0). On the other hand, because the dimension of the Cantor set is less than 1, it is *much* smaller than a one-dimensional set, so it is no surprise that its measure is zero. (This is a bit like saying that a surface has no volume, but now the two dimensions are 0.63 and 1 instead of 2 and 3.)

The most useful theory of fractional dimension is one developed by HAUSDORFF [VI.68]. One begins with a concept known as Hausdorff measure, which is a natural way of assessing the “ d -dimensional volume” of a set, even if d is not an integer. Suppose you have a curve in \mathbb{R}^3 and you want to work out its length by considering how easy it is to cover it with spheres. A first idea might be to say that the length was the smallest you could make the sum of the diameters of the spheres. But this does not work: you might be lucky and find that a long curve was tightly wrapped up, in which case you could cover it with a single sphere of small diameter.

However, this would no longer be possible if your spheres were required to be small. Suppose, therefore, that we require all the diameters of the spheres to be at most δ . Let $L(\delta)$ be the smallest we can then get the sum of the diameters to be. The smaller δ is, the less flexibility we have, so the larger $L(\delta)$ will be. Therefore, $L(\delta)$ tends to a (possibly infinite) limit L as δ tends to 0, and we call L the length of the curve.

Now suppose that we have a smooth surface in \mathbb{R}^3 and want to deduce its area from information about covering it with spheres. This time, the area that you can cover with a very small sphere (so small that it meets only one portion of the surface and that portion is almost flat) will be roughly proportional to the *square* of the diameter of the sphere. But that is the only

detail we need to change: let $A(\delta)$ be the smallest we can make the sum of the squares of the diameters of a set of spheres that cover the surface, if all those spheres have diameter at most δ . Then declare the area of the surface to be the limit of $A(\delta)$ as δ tends to 0. (Strictly speaking, we ought to multiply this limit by $\pi/4$, but then we get a definition that does not generalize easily.)

We have just given a way of defining length and area, for shapes in \mathbb{R}^3 . The only difference between the two was that for length we considered the sum of the diameters of small spheres, while for area we considered the sum of the *squares* of the diameters of small spheres. In general, we define the *d -dimensional Hausdorff measure* in a similar way, but considering the sum of the d th powers of the diameters.

We can use the concept of Hausdorff measure to give a rigorous definition of fractional dimension. It is not hard to show that for any shape X there will be exactly one appropriate d , in the following sense: if c is less than d , then the c -dimensional Hausdorff measure of X is 0, while if c is greater than d , then it is infinite. (For instance, the c -dimensional Hausdorff measure of a smooth surface is 0 if $c < 2$ and infinite if $c > 2$.) This d is called the *Hausdorff dimension* of the set X . Hausdorff dimension is very useful for analyzing fractal sets, which are discussed further in DYNAMICS [IV.15].

It is important to realize that the Hausdorff dimension of a set need not equal its topological dimension. For example, the Cantor set has topological dimension zero and Hausdorff dimension $\log 2 / \log 3$. A larger example is a very wiggly curve known as the *Koch snowflake*. Because it is a curve (and a single point is enough to cut it into two) it has topological dimension 1. However, because it is very wiggly, it has infinite length, and its Hausdorff dimension is in fact $\log 4 / \log 3$.

III.18 Distributions

Terence Tao

A function is normally defined to be an object $f : X \rightarrow Y$ which assigns to each point x in a set X , known as the *domain*, a point $f(x)$ in another set Y , known as the *range* (see THE LANGUAGE AND GRAMMAR OF MATHEMATICS [I.2 §2.2]). Thus, the definition of functions is set-theoretic and the fundamental operation that one can perform on a function is *evaluation*: given an element x of X , one evaluates f at x to obtain the element $f(x)$ of Y .

However, there are some fields of mathematics where this may not be the best way of describing functions. In geometry, for instance, the fundamental property of a function is not necessarily how it acts on points, but rather how it *pushes forward* or *pulls back* objects that are more complicated than points (e.g., other functions, BUNDLES [IV.10 §5] and sections, SCHEMES [IV.6 §3] and sheaves, etc.). Similarly, in analysis, a function need not necessarily be defined by what it does to points, but may instead be defined by what it does to objects of different kinds, such as sets or other functions; the former leads to the notion of a *measure*; the latter to that of a *distribution*.

Of course, all these notions of function and function-like objects are related. In analysis, it is helpful to think of the various notions of a function as forming a spectrum, with very “smooth” classes of functions at one end and very “rough” ones at the other. The smooth classes of functions are very restrictive in their membership: this means that they have good properties, and there are many operations that one can perform on them (such as, for example, differentiation), but it also means that one cannot necessarily ensure that the functions one is working with belong to this category. Conversely, the rough classes of functions are very general and inclusive: it is easy to ensure that one is working with them, but the price one pays is that the number of operations one can perform on these functions is often sharply reduced (see FUNCTION SPACES [III.29]).

Nevertheless, the various classes of functions can often be treated in a unified manner, because it is often possible to approximate rough functions arbitrarily well (in an appropriate TOPOLOGY [III.92]) by smooth ones. Then, given an operation that is naturally defined for smooth functions, there is a good chance that there will be exactly one natural way to extend it to an operation on rough functions: one takes a sequence of better and better smooth approximations to the rough functions, performs the operation on them, and passes to the limit.

Distributions, or *generalized functions*, belong at the rough end of the spectrum, but before we say what they are, it will be helpful to begin by considering some smoother classes of functions, partly for comparison and partly because one obtains rough classes of functions from smooth ones by a process known as *duality*: a *linear functional* defined on a space E of functions is simply a linear map ϕ from E to the scalars \mathbb{R} or \mathbb{C} . Typically, E is a normed space, or at least comes with a

topology, and the *dual space* is the space of *continuous* linear functionals.

The class $C^\omega[-1, 1]$ of analytic functions. These are in many ways the “nicest” functions of all, and include many familiar functions such as $\exp(x)$, $\sin(x)$, polynomials, and so on. However, we shall not discuss them further, because for many purposes they form too rigid a class to be useful. (For example, if an analytic function is zero everywhere on an interval, then it is forced to be zero everywhere.)

PUP query:
paragraph heading
OK here? Make
bold heading same
size as
surrounding text
perhaps?

The class $C_c^\infty[-1, 1]$ of test functions. These are the smooth (that is, infinitely differentiable) functions f , defined on the interval $[-1, 1]$, that vanish on neighborhoods of 1 and -1 . (That is, one can find $\delta > 0$ such that $f(x) = 0$ whenever $x > 1 - \delta$ or $x < -1 + \delta$.) They are more numerous than analytic functions and therefore more tractable for analysis. For instance, it is often useful to construct smooth “cutoff functions,” which are functions that vanish outside some small set but do not vanish inside it. Also, all the operations from calculus (differentiation, integration, composition, convolution, evaluation, etc.) are available for these functions.

The class $C^0[-1, 1]$ of continuous functions. These functions are regular enough for the notion of evaluation, $x \mapsto f(x)$, to make sense for every $x \in [-1, 1]$, and one can integrate such functions and perform algebraic operations such as multiplication and composition, but they are not regular enough that operations such as differentiation can be performed on them. Still, they are usually considered among the smoother examples of functions in analysis.

The class $L^2[-1, 1]$ of square-integrable functions. These are measurable functions $f : [-1, 1] \rightarrow \mathbb{R}$ for which the Lebesgue integral $\int_{-1}^1 |f(x)|^2 dx$ is finite. Usually one regards two such functions f and g as equal if the set of x such that $f(x) \neq g(x)$ has measure zero. (Thus, from the set-theoretic point of view, the object in question is really an EQUIVALENCE CLASS [I.2 §2.3] of functions.) Since a singleton $\{x\}$ has measure zero, we can change the value of $f(x)$ without changing the function. Thus, the notion of evaluation does not make sense for a square-integrable function $f(x)$ at any specific point x . However, two functions that differ on a set of measure zero have the same LEBESGUE INTEGRAL [III.57], so integration does make sense.

A key point about this class is that it is *self-dual* in the following sense. Any two functions in this class can be paired together by the *inner product*

$\langle f, g \rangle = \int_{-1}^1 f(x)g(x) dx$. Therefore, given a function $g \in L^2[-1, 1]$, the map $f \mapsto \langle f, g \rangle$ defines a linear functional on $L^2[-1, 1]$, which turns out to be continuous. Moreover, given any continuous linear functional ϕ on $L^2[-1, 1]$, there is a unique function $g \in L^2[-1, 1]$ such that $\phi(f) = \langle f, g \rangle$ for every f . This is a special case of one of the *Riesz representation theorems*.

The class $C^0[-1, 1]^*$ of finite Borel measures. Any finite Borel MEASURE [III.57] μ gives rise to a continuous linear functional on $C^0[-1, 1]$ defined by $f \mapsto \langle \mu, f \rangle = \int_{-1}^1 f(x) d\mu$. Another of the Riesz representation theorems says that every continuous linear functional on $C^0[-1, 1]$ arises in this way, so one could in principle define a finite Borel measure to be a continuous linear functional on $C^0[-1, 1]$.

The class $C^\infty([-1, 1])^*$ of distributions. Just as measures can be viewed as continuous linear functionals on $C^0([-1, 1])$, a *distribution* μ is a continuous linear functional on $C_c^\infty([-1, 1])$ (with an appropriate topology). Thus, a distribution can be viewed as a “virtual function”: it cannot itself be directly evaluated, or even integrated over an open set, but it can still be paired with any test function $g \in C_c^\infty([-1, 1])$, producing a number $\langle \mu, g \rangle$. A famous example is the *Dirac distribution* δ_0 , defined as the functional which, when paired with any test function g , returns the evaluation $g(0)$ of g at zero: $\langle \delta_0, g \rangle = g(0)$. Similarly, we have the derivative of the Dirac distribution, $-\delta'_0$, which, when paired with any test function g , returns the derivative $g'(0)$ of g at zero: $\langle -\delta'_0, g \rangle = g'(0)$. (The reason for the minus sign will be given later.) Since test functions have so many operations available to them, there are many ways to define continuous linear functionals, so the class of distributions is quite large. Despite this, and despite the indirect, virtual nature of distributions, one can still define many operations on them; we shall discuss this later.

The class $C^\omega([-1, 1])^*$ of hyperfunctions. There are classes of functions more general still than distributions. For instance, there are hyperfunctions, which roughly speaking one can think of as linear functionals that can be tested only against analytic functions $g \in C^\omega([-1, 1])$ rather than against test functions $g \in C_c^\infty([-1, 1])$. However, as the class of analytic functions is so sparse, hyperfunctions tend not to be as useful in analysis as distributions.

At first glance, the concept of a distribution has limited utility, since all a distribution μ is empowered to do

is to be tested against test functions g to produce inner products $\langle \mu, g \rangle$. However, using this inner product, one can often take operations that are initially defined only on test functions, and *extend* them to distributions by duality. A typical example is differentiation. Suppose one wants to know how to define the derivative μ' of a distribution, or in other words how to define $\langle \mu', g \rangle$ for any test function g and distribution μ . If μ is itself a test function $\mu = f$, then we can evaluate this using integration by parts (recalling that test functions vanish at -1 and 1). We have

$$\begin{aligned} \langle f', g \rangle &= \int_{-1}^1 f'(x)g(x) dx \\ &= - \int_{-1}^1 f(x)g'(x) dx = -\langle f, g' \rangle. \end{aligned}$$

Note that if g is a test function, then so is g' . We can therefore generalize this formula to arbitrary distributions by defining $\langle \mu', g \rangle = -\langle \mu, g' \rangle$. This is the justification for the differentiation of the Dirac distribution: $\langle \delta'_0, g \rangle = -\langle \delta_0, g' \rangle = -g'(0)$.

More formally, what we have done here is to compute the adjoint of the differentiation operation (as defined on the dense space of test functions). Then we have taken adjoints again to define the differentiation operation for general distributions. This procedure is well-defined and also works for many other concepts; for instance, one can add two distributions, multiply a distribution by a smooth function, convolve two distributions, and compose distributions on both left and right with suitably smooth functions. One can even take Fourier transforms of distributions. For instance, the Fourier transform of the Dirac delta δ_0 is the constant function 1, and vice versa (this is essentially the Fourier inversion formula), while the distribution $\sum_{n \in \mathbb{Z}} \delta_0(x - n)$ is its own Fourier transform (this is essentially the Poisson summation formula). Thus the space of distributions is quite a good space to work in, in that it contains a large class of functions (e.g., all measures and integrable functions), and is also closed under a large number of common operations in analysis. Because the test functions are dense in the space of distributions, the operations as defined on distributions are usually compatible with those on test functions. For instance, if f and g are test functions and $f' = g$ in the sense of distributions, then $f' = g$ will also be true in the classical sense. This often allows one to manipulate distributions as if they were test functions without fear of confusion or inaccuracy. The main operations one has to be careful about are evaluation and pointwise multiplication of distributions, both

T&T note: this paragraph can be cut, according to Terry and Tim, if space becomes really tight.

of which are usually not well-defined (e.g., the square of the Dirac delta distribution is not well-defined as a distribution).

Another way to view distributions is as the *weak limit* of test functions. A sequence of functions f_n is said to *converge weakly* to a distribution μ if $\langle f_n, g \rangle \rightarrow \langle \mu, g \rangle$ for all test functions g . For instance, if φ is a test function with total integral $\int_{-1}^1 \varphi = 1$, then the test functions $f_n(x) = n\varphi(nx)$ can be shown to converge weakly to the Dirac delta distribution δ_0 , while the functions $f'_n = n^2\varphi'(nx)$ converge weakly to the derivative δ'_0 of the Dirac delta. On the other hand, the functions $g_n(x) = \cos(nx)\varphi(x)$ converge weakly to zero (this is a variant of the *Riemann-Lebesgue lemma*). Thus weak convergence has some unusual features not present in stronger notions of convergence, in that severe oscillations can sometimes “disappear” in the limit. One advantage of working with distributions instead of smoother functions is that one often has some compactness in the space of distributions under weak limits (e.g., by the Banach-Alaoglu theorem). Thus, distributions can be thought of as asymptotic extremes of behavior of smoother functions, just as real numbers can be thought of as limits of rational numbers.

Because distributions can be easily differentiated, while still being closely connected to smoother functions, they have been extremely useful in the study of partial differential equations (PDEs), particularly when the equations are linear. For instance, the general solution to a linear PDE can often be described in terms of its *fundamental solution*, which solves the PDE in the sense of distributions. More generally, distribution theory (together with related concepts, such as that of a *weak derivative*) gives an important (though certainly not the only) means to define *generalized solutions* of both linear and nonlinear PDEs. As the name suggests, these generalize the concept of smooth (or *classical*) solutions by allowing the formation of singularities, shocks, and other nonsmooth behavior. In some cases the easiest way to construct a smooth solution to a PDE is first to construct a generalized solution and then to use additional arguments to show that the generalized solution is in fact smooth.

III.19 Duality

Duality is an important general theme that has manifestations in almost every area of mathematics. Over and over again, it turns out that one can associate with a given mathematical object a related, “dual” object that

helps one to understand the properties of the object one started with. Despite the importance of duality in mathematics, there is no single definition that covers all instances of the phenomenon. So let us look at a few examples and at some of the characteristic features that they exhibit.

1 Platonic Solids

Suppose you take a cube, draw points at the centers of each of its six faces, and let those points be the vertices of a new polyhedron. The polyhedron you get will be a regular octahedron. What happens if you repeat the process? If you draw a point at the center of each of the eight faces of the octahedron, you will find that these points are the eight vertices of a cube. We say that the cube and the octahedron are *dual* to one another. The same can be done for the other Platonic solids: the dodecahedron and the icosahedron are dual to one another, while the dual of a tetrahedron is again a tetrahedron.

The duality just described does more than just split up the five Platonic solids into three groups: it allows us to associate statements about a solid with statements about its dual. For instance, two faces of a dodecahedron are *adjacent* if they share an edge, and this is so if and only if the corresponding vertices of the dual icosahedron are linked by an edge. And for this reason there is also a correspondence between edges of the dodecahedron and edges of the icosahedron.

2 Points and Lines in the Projective Plane

There are several equivalent definitions of the PROJECTIVE PLANE [I.3 §6.7]. One, which we shall use here, is that it is the set of all lines in \mathbb{R}^3 that go through the origin. These lines we call the “points” of the projective plane. In order to visualize this set as a geometrical object and to make its “points” more point-like, it is helpful to associate each line through the origin with the pair of points in \mathbb{R}^3 at which it intersects the unit sphere: indeed, one can define the projective plane as the unit sphere with opposite points identified.

A typical “line” in the projective plane is the set of all “points” (that is, lines through the origin) that lie in some plane through the origin. This is associated with the great circle in which that plane intersects the unit sphere, once again with opposite points identified.

There is a natural association between lines and points in the projective plane: each point P is associated with the line L that consists of all points orthogonal to

P, and each line L is associated with the single point P that is orthogonal to all points in L. For example, if P is the z-axis, then the associated projective line L is the set of all lines through the origin that lie in the xy -plane, and vice versa. This association has the following basic property: if a point P belongs to a line L, then the line associated with P contains the point associated with L.

This allows us to translate statements about points and lines into logically equivalent statements about lines and points. For example, three points are collinear (that is, they all lie in a line) if and only if the corresponding lines are concurrent (that is, there is some point that is contained in all of them). In general, once you have proved a theorem in projective geometry, you get another, dual, theorem for free (unless the dual theorem turns out to be the same as the original one).

3 Sets and Their Complements

Let X be a set. If A is any subset of X , then the *complement* of A , written A^c , is the set of all elements of X that do not belong to A . The complement of the complement of A is clearly A , so there is a kind of duality between sets and their complements. *De Morgan's laws* are the statements that $(A \cap B)^c = A^c \cup B^c$ and $(A \cup B)^c = A^c \cap B^c$: they tell us that complementation “turns intersections into unions,” and vice versa. Notice that if we apply the first law to A^c and B^c , then we find that $(A^c \cap B^c)^c = A \cup B$. Taking complements of both sides of this equality gives us the second law.

Because of de Morgan's laws, any identity involving unions and intersections remains true when you interchange them. For example, one useful identity is $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$. Applying this to the complements of the sets and using de Morgan's laws, it is straightforward to deduce the equally useful identity $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

4 Dual Vector Spaces

Let V be a VECTOR SPACE [I.3 §2.3], over \mathbb{R} , say. The *dual space* V^* is defined to be the set of all *linear functionals* on V : that is, linear maps from V to \mathbb{R} . It is not hard to define appropriate notions of addition and scalar multiplication and show that these make V^* into a vector space as well.

Suppose that T is a LINEAR MAP [I.3 §4.2] from a vector space V to a vector space W . If we are given an element w^* of the dual space W^* , then we can use T and w^* to create an element of V^* as follows: it is the map that takes v to the real number $w^*(Tv)$. This map,

which is denoted by T^*w^* , is easily checked to be linear. The function T^* is itself a linear map, called the *adjoint* of T , and it takes elements of W^* to elements of V^* .

This is a typical feature of duality: a function f from object A to object B very often gives rise to a function g from the dual of B to the dual of A .

Suppose that T^* is a surjection. Then if $v \neq v'$, we can find v^* such that $v^*(v) \neq v^*(v')$, and then $w^* \in W^*$ such that $T^*w^* = v^*$, so that $T^*w^*(v) \neq T^*w^*(v')$, and hence $w^*(Tv) \neq w^*(Tv')$. This implies that $Tv \neq Tv'$, which proves that T is an injection. We can also prove that if T^* is an injection, then T is a surjection. Indeed, if T is not a surjection, then TV is a proper subspace of W , which allows us to find a nonzero linear functional w^* such that $w^*(Tv) = 0$ for every $v \in V$, and hence such that $T^*w^* = 0$, which contradicts the injectivity of T^* . If V and W are finite dimensional, then $(T^*)^* = T$, so in this case we find that T is an injection if and only if T^* is a surjection, and vice versa. Therefore, we can use duality to convert an existence problem into a uniqueness problem. This conversion of one kind of problem into a different kind is another characteristic and very useful feature of duality.

If a vector space has additional structure, the definition of the dual space may well change. For instance, if X is a real BANACH SPACE [III.64], then X^* is defined to be the space of all *continuous* linear functionals from X to \mathbb{R} , rather than the space of *all* linear functionals. This space is also a Banach space: the norm of a continuous linear functional f is defined to be $\sup\{|f(x)| : x \in X, \|x\| \leq 1\}$. If X is an explicit example of a Banach space (such as one of the spaces discussed in FUNCTION SPACES [III.29]), it can be extremely useful to have an explicit description of the dual space. That is, one would like to find an explicitly described Banach space Y and a way of associating with each nonzero element y of Y a nonzero continuous linear functional ϕ_y defined on X , in such a way that every continuous linear functional is equal to ϕ_y for some $y \in Y$.

From this perspective, it is more natural to regard X and Y as having the same status. We can reflect this in our notation by writing $\langle x, y \rangle$ instead of $\phi_y(x)$. If we do this, then we are drawing attention to the fact that the map $\langle \cdot, \cdot \rangle$, which takes the pair (x, y) to the real number $\langle x, y \rangle$, is a continuous bilinear map from $X \times Y$ to \mathbb{R} .

More generally, whenever we have two mathematical objects A and B , a set S of “scalars” of some kind, and a function $\beta : A \times B \rightarrow S$ that is a structure-preserving map in each variable separately, we can think of the elements of A as elements of the dual of B , and vice versa. Functions like β are called *pairings*.

5 Polar Bodies

Let X be a subset of \mathbb{R}^n and let $\langle \cdot, \cdot \rangle$ be the standard INNER PRODUCT [III.37] on \mathbb{R}^n . Then the *polar* of X , denoted X° , is the set of all points $y \in \mathbb{R}^n$ such that $\langle x, y \rangle \leq 1$ for every $x \in X$. It is not hard to check that X° is closed and convex, and that if X is closed and convex, then $(X^\circ)^\circ = X$. Furthermore, if $n = 3$ and X is a Platonic solid centered at the origin, then X° is (a multiple of) the dual Platonic solid, and if X is the “unit ball” of a normed space (that is, the set of all points of norm at most 1), then X° is (easily identified with) the unit ball of the dual space.

6 Duals of Abelian Groups

If G is an Abelian group, then a *character* on G is a homomorphism from G to the group \mathbb{T} of all complex numbers of modulus 1. Two characters can be multiplied together in an obvious way, and this multiplication makes the set of all characters on G into another Abelian group, called the *dual group*, \hat{G} , of the group G . Again, if G has a topological structure, then one usually imposes an additional continuity condition.

An important example is when the group is itself \mathbb{T} . It is not hard to show that the continuous homomorphisms from \mathbb{T} to \mathbb{T} all have the form $e^{i\theta} \mapsto e^{in\theta}$ for some integer n (which may be negative or zero). Thus, the dual of \mathbb{T} is (isomorphic to) \mathbb{Z} .

This form of duality between groups is called *Pontryagin duality*. Note that there is an easily defined pairing between G and \hat{G} : given an element $g \in G$ and a character $\psi \in \hat{G}$, we define $\langle g, \psi \rangle$ to be $\psi(g)$.

Under suitable conditions, this pairing extends to *functions* defined on G and \hat{G} . For instance, if G and \hat{G} are finite, and $f : G \rightarrow \mathbb{C}$ and $F : \hat{G} \rightarrow \mathbb{C}$, then we can define $\langle f, F \rangle$ to be the complex number $|G|^{-1} \sum_{g \in G} \sum_{\psi \in \hat{G}} f(g) F(\psi)$. In general, one obtains a pairing between a complex HILBERT SPACE [III.37] of functions on G and a Hilbert space of functions on \hat{G} .

This extended pairing leads to another important duality. Given a function in the Hilbert space $L^2(\mathbb{T})$, its *Fourier transform* is the function $\hat{f} \in \ell_2(\mathbb{Z})$ that

is defined by the formula

$$\hat{f}(n) = \frac{1}{2\pi} \int_0^{2\pi} f(e^{i\theta}) e^{-in\theta} d\theta.$$

The Fourier transform, which can be defined similarly for functions on other Abelian groups, is immensely useful in many areas of mathematics. (See, for example, FOURIER TRANSFORMS [III.27] and REPRESENTATION THEORY [IV.12].) By contrast with some of the previous examples, it is *not* always easy to translate a statement about a function f into an equivalent statement about its Fourier transform \hat{f} , but this is what gives the Fourier transform its power: if you wish to understand a function f defined on \mathbb{T} , then you can explore its properties by looking at both f and \hat{f} . Some properties will follow from facts that are naturally expressed in terms of f and others from facts that are naturally expressed in terms of \hat{f} . Thus, the Fourier transform “doubles one’s mathematical power.”

7 Homology and Cohomology

Let X be a compact n -dimensional MANIFOLD [I.3 §6.9]. If M and M' are an i -dimensional submanifold and an $(n - i)$ -dimensional submanifold of X , respectively, and if they are well-behaved and in sufficiently general position, then they will intersect in a finite set of points. If one assigns either 1 or -1 to each of these points in a natural way that takes account of how M and M' intersect, then the sum of the numbers at the points is an invariant called the *intersection number* of M and M' . This number turns out to depend only on the HOMOLOGY CLASSES [IV.10 §4] of M and M' . Thus, it defines a map from $H_i(X) \times H_{n-i}(X)$ to \mathbb{Z} , where we write $H_r(X)$ for the r th homology group of X . This map is a group homomorphism in each variable separately, and the resulting pairing leads to a notion of duality called *Poincaré duality*, and ultimately to the modern theory of *cohomology*, which is dual to homology. As with some of our other examples, many concepts associated with homology have dual concepts: for example, in homology one has a *boundary map*, whereas in cohomology there is a *coboundary map* (in the opposite direction). Another example is that a continuous map from X to Y gives rise to a homomorphism from the homology group $H_i(X)$ to the homology group $H_i(Y)$, and also to a homomorphism from the cohomology group $H^i(Y)$ to the cohomology group $H^i(X)$.

8 Further Examples Discussed in This Book

The examples above are not even close to a complete list: even in this book there are several more. For instance, the article on DIFFERENTIAL FORMS [III.16] discusses a pairing, and hence a duality, between k -forms and k -dimensional surfaces. (The pairing is given by integrating the form over the surface.) The article on DISTRIBUTIONS [III.18] shows how to use duality to give rigorous definitions of function-like objects such as the Dirac delta function. The article on MIRROR SYMMETRY [IV.14] discusses an astonishing (and still largely conjectural) duality between CALABI-YAU MANIFOLDS [III.6] and so-called “mirror manifolds.” Often the mirror manifold is much easier to understand than the original manifold, so this duality, like the Fourier transform, makes certain calculations possible that would otherwise be unthinkable. And the article on REPRESENTATION THEORY [IV.12] discusses the “Langlands dual” of certain (non-Abelian) groups: a proper understanding of this duality would solve many major open problems.

III.20 Dynamical Systems and Chaos

From a scientific point of view, a dynamical system is a physical system, such as a collection of planets or the water in a canal, that changes over time. Typically, the positions and velocities of the parts of such a system at a time t depend only on the positions and velocities of those parts just before that time, which means that the behavior of the system is governed by a system of PARTIAL DIFFERENTIAL EQUATIONS [I.3 §5.3]. Often, a very simple collection of partial differential equations can lead to very complicated behavior of the physical system.

From a mathematical point of view, a dynamical system is any mathematical object that evolves in time according to a precise rule that determines the behavior of the system at time t from its behavior just beforehand. Sometimes, as above, “just beforehand” refers to a time infinitesimally earlier, which is why calculus is involved. But there is also a vigorous theory of *discrete* dynamical systems, where the “time” t takes integer values, and the “time just before t ” is $t - 1$. If f is the function that tells us how the system at time t depends on the system at time $t - 1$, then the system as a whole can be thought of as the process of *iterating* f : that is, applying f over and over again.

As with continuous dynamical systems, a very simple function f can lead to very complicated behavior if you iterate it enough times. In particular, some of the most interesting dynamical systems, both discrete ones and continuous ones, exhibit an extreme sensitivity to initial conditions, which is known as *chaos*. This is true, for example, of the equations that govern weather. One cannot hope to specify exactly the wind speed at every point on the Earth’s surface (not to mention high above it), which means that one has to make do with approximations. Because the relevant equations are chaotic, the resulting inaccuracies, which may be small to start with, rapidly propagate and overwhelm the system: you could start with a different, equally good approximation and find that after a fairly short time the system had evolved in a completely different way. This is why accurate forecasting is impossible more than a few days in advance.

For more about dynamical systems and chaos, see DYNAMICS [IV.15].

III.21 Elliptic Curves

Jordan S. Ellenberg

An elliptic curve over a field K can be defined as an algebraic curve of genus 1 over K , endowed with a point defined over K . If this definition is too abstract for your tastes, then an equivalent definition is the following: an elliptic curve is a curve in the plane determined by an equation of the form

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6. \quad (1)$$

When the characteristic of K is not 2, we can transform this equation into the simpler form $y^2 = f(x)$, for some cubic polynomial f . In this sense, an elliptic curve is a rather concrete object. However, this definition has given rise to a subject of seemingly inexhaustible mathematical interest, which has provided a tremendous fund of ideas, examples, and problems in number theory and algebraic geometry. This is in part because there are many values of “ X ” for which it is the case that “the simplest interesting example of X is an elliptic curve.”

For instance, the points of an elliptic curve E with coordinates in K naturally form an Abelian group, which we call $E(K)$. The connected projective VARIETIES [III.97] that admit a group law of this kind are called *Abelian varieties*; and elliptic curves are just the Abelian varieties that are one dimensional. The

Mordell-Weil theorem tells us that, when K is a number field and A is an Abelian variety, $A(K)$ is actually a *finitely generated* Abelian group, called a *Mordell-Weil group*; these Abelian groups are much studied but have retained much of their mystery (see RATIONAL POINTS ON CURVES AND THE MORDELL CONJECTURE [V.31]). Even when A is an elliptic curve, in which case we would call it E instead, there is a great deal that we do not know, though THE BIRCH-SWINNERTON-DYER CONJECTURE [V.4] offers a conjectural formula for the rank of the group $E(K)$. For much more on the topic of rational points on elliptic curves, see ARITHMETIC GEOMETRY [IV.6].

Since $E(K)$ forms an Abelian group, given any prime p one can look at the subgroup of elements P such that $pP = 0$. This subgroup is called $E(K)[p]$. In particular, we can take the algebraic closure \bar{K} of K and look at $E(\bar{K})[p]$. It turns out that, when K is a NUMBER FIELD [III.65] (or, for that matter, any field of characteristic not equal to p), this group is isomorphic to $(\mathbb{Z}/p\mathbb{Z})^2$, no matter what choice of E we started with. If the group is the same for all elliptic curves, why is it interesting? Because it turns out that the GALOIS GROUP [V.24] $\text{Gal}(\bar{K}/K)$ permutes the set $E(\bar{K})[p]$. In fact, the action of $\text{Gal}(\bar{K}/K)$ on the group $(\mathbb{Z}/p\mathbb{Z})^2$ gives rise to a REPRESENTATION [III.79] of the Galois group. This is a foundational example in the theory of *Galois representations*, which has become central to contemporary number theory. Indeed, the proof of FERMAT'S LAST THEOREM [V.12] by Andrew Wiles is in the end a theorem about the Galois representations that arise from elliptic curves. And what Wiles proved about these special Galois representations is itself a small special case of the family of conjectures known as the *Langlands program*, which proposes a thoroughgoing correspondence between Galois representations and *automorphic forms*, which are generalized versions of the classical analytic functions called MODULAR FORMS [III.61].

In another direction, if E is an elliptic curve over \mathbb{C} , then the set of points of E with complex coordinates, which we denote $E(\mathbb{C})$, is a COMPLEX MANIFOLD [III.90 §3]. It turns out that this manifold can always be expressed as the quotient of the complex plane by a certain group Λ of transformations. What is more, these transformations are just translations: each map sends z to $z + c$ for some complex number c . (This expression of $E(\mathbb{C})$ as a quotient is carried out with the help of ELLIPTIC FUNCTIONS [V.34].) Each elliptic curve gives rise in this way to a subset—indeed, a subgroup—of the complex numbers; the elements of this subgroup

are called *periods* of the elliptic curve. This construction can be regarded as the very beginning of *Hodge theory*, a powerful branch of algebraic geometry with a reputation for extreme difficulty. (The *Hodge conjecture*, a central question in the theory, is one of the Clay Institute's million-dollar-prize problems.)

Yet another point of view is presented by the MODULI SPACE [IV.8] of elliptic curves, denoted $M_{1,1}$. This is itself a curve, but not an elliptic one. (In fact, if I am completely honest, I should say that $M_{1,1}$ is not quite a curve at all—it is an object called, depending on whom you ask, an ORBIFOLD [IV.7 §7] or an *algebraic stack*—you can think of it as a curve from which someone has removed a few points, folded the points in half or into thirds, and then glued the folded-up points back in. You might find it reassuring to know that even professionals in the subject find this process rather difficult to visualize.) The curve $M_{1,1}$ is a “simplest example” in two ways: it is the simplest *modular curve*, and simultaneously the simplest moduli space of curves.

III.22 The Euclidean Algorithm and Continued Fractions

Keith Ball

1 The Euclidean Algorithm

THE FUNDAMENTAL THEOREM OF ARITHMETIC [V.16], which states that every integer can be factored into primes in a unique way, has been known since antiquity. The usual proof depends upon what is known as the Euclidean algorithm, which constructs the highest common factor (h , say) of two numbers m and n . In doing so, it shows that h can be written in the form $am + bn$ for some pair of integers a, b (not necessarily positive). For example, the highest common factor of 17 and 7 is 1, and sure enough we can express 1 as the combination $1 = 5 \times 17 - 12 \times 7$.

The algorithm works as follows. Assume that m is larger than n and start by dividing m by n to yield a quotient q_1 and a remainder r_1 that is less than n . Then we have

$$m = q_1 n + r_1. \quad (1)$$

Now since $r_1 < n$ we may divide n by r_1 to obtain a second quotient and remainder:

$$n = q_2 r_1 + r_2. \quad (2)$$

Continue in this way, dividing r_1 by r_2 , r_2 by r_3 , and so on. The remainders get smaller each time but cannot go below zero. So the process must stop at some point

with a remainder of 0: that is, with a division that comes out exactly. For instance, if $m = 165$ and $n = 70$, the algorithm generates the sequence of divisions

$$165 = 2 \times 70 + 25, \quad (3)$$

$$70 = 2 \times 25 + 20, \quad (4)$$

$$25 = 1 \times 20 + 5, \quad (5)$$

$$20 = 4 \times 5 + 0. \quad (6)$$

The process guarantees that the last nonzero remainder, 5 in this case, is the highest common factor of m and n . On the one hand, the last line shows that 5 is a factor of the previous remainder 20. Now the last-but-one line shows that 5 is also a factor of the remainder 25 that occurred one step earlier, because 25 is expressed as a combination of 20 and 5. Working back up the algorithm we conclude that 5 is a factor of both $m = 165$ and $n = 70$. So 5 is certainly a common factor of m and n .

On the other hand, the last-but-one line shows that 5 can be written as a combination of 25 and 20 with integer coefficients. Since the previous line shows that 20 can be written as a combination of 70 and 25 we can write 5 in terms of 70 and 25:

$$5 = 25 - 20 = 25 - (70 - 2 \times 25) = 3 \times 25 - 70.$$

Continuing back up the algorithm we can express 25 in terms of 165 and 70 and conclude that

$$5 = 3 \times (165 - 2 \times 70) - 70 = 3 \times 165 - 7 \times 70.$$

This shows that 5 is the *highest* common factor of 165 and 70 because any factor of 165 and 70 would automatically be a factor of $3 \times 165 - 7 \times 70$: that is, a factor of 5. Along the way we have shown that the highest common factor can be expressed as a combination of the two original numbers m and n .

2 Continued Fractions for Numbers

During the 1500 years following Euclid, it was realized by mathematicians of the Indian and Arabic schools that the application of the Euclidean algorithm to a pair of integers m and n could be encoded in a formula for the ratio m/n . The equation (1) can be written

$$\frac{m}{n} = q_1 + \frac{r_1}{n} = q_1 + \frac{1}{F},$$

where $F = n/r_1$. Now equation (2) expresses F as

$$F = q_2 + \frac{r_2}{r_1}.$$

The next step of the algorithm will produce an expression for r_1/r_2 and so on. If the algorithm stops after

k steps, then we can put these expressions together to get what is called the *continued fraction* for m/n :

$$\frac{m}{n} = q_1 + \frac{1}{q_2 + \frac{1}{q_3 + \frac{1}{\ddots + \frac{1}{q_k}}}}.$$

For example,

$$\frac{165}{70} = 2 + \frac{1}{2 + \frac{1}{1 + \frac{1}{4}}}.$$

The continued fraction can be constructed directly from the ratio $165/70 = 2.35714\dots$ without reference to the integers 165 and 70. We start by subtracting from 2.35714... the largest whole number we can: namely 2. Now we take the reciprocal of what is left: $1/0.35714\dots = 2.8$. Again we subtract off the largest integer we can, 2, which tells us that $q_2 = 2$. The reciprocal of 0.8 is 1.25, so $q_3 = 1$ and then, finally, $1/0.25 = 4$, so $q_4 = 4$ and the continued fraction stops.

The mathematician John Wallis, who worked in the seventeenth century, seems to have been the first to give a systematic account of continued fractions and to recognize that continued-fraction expansions exist for all numbers (not only rational numbers), provided that we allow the continued fraction to have infinitely many levels. If we start with any positive number, we can build its continued fraction in the same way as for the ratio 2.35714.... For example, if the number is $\pi = 3.14159265\dots$, we start by subtracting 3, then take the reciprocal of what is left: $1/0.14159\dots = 7.06251\dots$. So for π we get that the second quotient is 7. Continuing the process we build the continued fraction

$$\pi = 3 + \frac{1}{7 + \frac{1}{15 + \frac{1}{1 + \frac{1}{292 + \frac{1}{1 + \frac{1}{\ddots}}}}}}. \quad (7)$$

The numbers 3, 7, 15, and so on, that appear in the fraction are called the *partial quotients* of π .

The continued fraction for a real number can be used to approximate it by rational numbers. If we truncate the continued fraction after several steps, we are left with a finite continued fraction which is a rational number: for example, by truncating the fraction (7), one level down we get the familiar approximation $\pi \approx 3 + 1/7 = 22/7$; at the second level we get the approximation $3 + 1/(7 + 1/15) = 333/106$. The truncations at different levels thus generate a sequence of rational approximations: the sequence for π begins

$$3, 22/7, 333/106, 355/113, \dots$$

Whatever positive number x we start with, the sequence of continued-fraction approximations will approach x as we move further down the fraction. Indeed, the formal interpretation of the equation (7) is precisely that the successive truncations of the fraction approach π .

Naturally, in order to get better approximations to a number x we need to take more “complicated” fractions—fractions with larger numerator and denominator. The continued-fraction approximations to x are *best* approximations to x in the following sense: if p/q is one of these fractions, then it is impossible to find any fraction r/s that is closer than p/q to x and that has denominator s smaller than q .

Moreover, if p/q is one of the approximations coming from the continued fraction for x , then the error $x - p/q$ cannot be too large relative to the size of the denominator q ; specifically, it is always true that

$$\left| x - \frac{p}{q} \right| \leq \frac{1}{q^2}. \quad (8)$$

This error estimate shows just how special the continued-fraction approximations are: if you pick a denominator q without thinking, and then select the numerator p that makes p/q closest to x , the only thing you can guarantee is that x lies between $(p - 1/2)/q$ and $(p + 1/2)/q$. So the error could be as large as $1/(2q)$, which is much bigger than $1/(q^2)$ if q is a large integer.

Sometimes a continued-fraction approximation to x can have even smaller error than is guaranteed by (8). For example, the approximation $\pi \approx 355/113$ that we get by truncating (7) at the third level is exceptionally accurate, the reason being that the next partial quotient, 292, is rather large. So we are not changing the fraction much by ignoring the tail $1/(292 + 1/(1 + \dots))$. In this sense, the most difficult number to approximate by fractions is the one with the smallest possible partial quotients, i.e., the one with all its partial quotients equal to 1. This number,

$$1 + \frac{1}{1 + \frac{1}{1 + \frac{1}{\ddots}}}, \quad (9)$$

can be easily calculated because the sequence of partial quotients is periodic: it repeats itself. If we call the number ϕ , then $\phi - 1$ is $1/(1 + 1/(1 + \dots))$. The reciprocal of this number is exactly the continued fraction (9) for ϕ . Hence

$$\frac{1}{\phi - 1} = \phi,$$

which in turn implies that $\phi^2 - \phi = 1$. The roots of this quadratic equation are $(1 + \sqrt{5})/2 = 1.618\dots$ and

$(1 - \sqrt{5})/2 = -0.618\dots$. Since the number we are trying to find is positive, it is the first of these roots: the so-called *golden ratio*.

It is quite easy to show that, just as (9) represents the positive solution of the equation $x^2 - x - 1 = 0$, any other periodic continued fraction represents a root of a quadratic equation. This fact seems to have been understood already in the sixteenth century. It is quite a lot trickier to prove the converse: that the continued fraction of any quadratic surd is periodic. This was established by LAGRANGE [VI.22] during the eighteenth century and is closely related to the existence of units in quadratic number fields (see ALGEBRAIC NUMBERS [IV.3]).

3 Continued Fractions for Functions

Several of the most important functions in mathematics are most easily described using infinite sums. For example, the EXPONENTIAL FUNCTION [III.25] has the infinite series

$$e^x = 1 + x + \frac{x^2}{2} + \dots + \frac{x^n}{n!} + \dots$$

There are also a number of functions that have simple continued-fraction expansions: continued fractions involving a variable like x . These are probably the most important continued fractions historically.

For example, the function $x \mapsto \tan x$ has the continued fraction

$$\tan x = \frac{x}{1 - \frac{x^2}{3 - \frac{x^2}{5 - \ddots}}}, \quad (10)$$

valid for any value of x other than the odd multiples of $\pi/2$, where the tangent function has a vertical asymptote.

Whereas the infinite series of a function can be truncated to provide *polynomial* approximations to the function, truncation of the continued fraction provides approximations by *rational functions*: functions that are ratios of polynomials. For instance, if we truncate the fraction for the tangent after one level, then we get the approximation

$$\tan x \approx \frac{x}{1 - x^2/3} = \frac{3x}{3 - x^2}.$$

This continued fraction, and the rapidity with which its truncations approach $\tan x$, played the central role in the proof that π is irrational: that π is not the ratio of two whole numbers. The proof was found by Johann Lambert in the 1760s. He used the continued fraction to show that if x is a rational number (other than 0),

then $\tan x$ is not. But $\tan \pi/4 = 1$ (which certainly is rational), so $\pi/4$ cannot be.

III.23 The Euler and Navier-Stokes Equations

Charles Fefferman

The Euler and Navier-Stokes equations describe the motion of an idealized fluid. They are important in science and engineering, yet they are very poorly understood. They present a major challenge to mathematics.

To state the equations we work in Euclidean space \mathbb{R}^d , with $d = 2$ or 3 . Suppose that, at position $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ and at time $t \in \mathbb{R}$, the fluid is moving with a velocity vector $u(x, t) = (u_1(x, t), \dots, u_d(x, t)) \in \mathbb{R}^d$, and the pressure in the fluid is $p(x, t) \in \mathbb{R}$. The Euler equation is

$$\left(\frac{\partial}{\partial t} + \sum_{j=1}^d u_j \frac{\partial}{\partial x_j} \right) u_i(x, t) = \frac{-\partial p}{\partial x_i}(x, t) \quad (i = 1, \dots, d) \quad (1)$$

for all (x, t) ; and the Navier-Stokes equation is

$$\begin{aligned} & \left(\frac{\partial}{\partial t} + \sum_{j=1}^d u_j \frac{\partial}{\partial x_j} \right) u_i(x, t) \\ &= \nu \left(\sum_{j=1}^d \frac{\partial^2}{\partial x_j^2} \right) u_i(x, t) - \frac{\partial p}{\partial x_i}(x, t) \quad (i = 1, \dots, d) \end{aligned} \quad (2)$$

for all (x, t) . Here, $\nu > 0$ is a coefficient of friction called the “viscosity” of the fluid.

In this article we restrict our attention to incompressible fluids, which means that, in addition to requiring that they satisfy (1) or (2), we also demand that

$$\operatorname{div} u \equiv \sum_{j=1}^d \frac{\partial u_j}{\partial x_j} = 0 \quad (3)$$

for all (x, t) . The Euler and Navier-Stokes equations are nothing but Newton’s law $F = ma$ applied to an infinitesimal portion of the fluid. In fact, the vector

$$\left(\frac{\partial}{\partial t} + \sum_{j=1}^d u_j \frac{\partial}{\partial x_j} \right) u$$

is easily seen to be the acceleration experienced by a molecule of fluid that finds itself at position x at time t .

The forces F leading to the Euler equation arise entirely from pressure gradients (e.g., if the pressure increases with height, then there is a net force pushing

the fluid down). The additional term

$$\nu \left(\sum_{j=1}^d \frac{\partial^2}{\partial x_j^2} \right) u$$

in (2) arises from frictional forces.

The Navier-Stokes equations agree very well with experiments on real fluids under many and varied circumstances. Since fluids are important, so are the Navier-Stokes equations.

The Euler equation is simply the limiting case $\nu = 0$ of Navier-Stokes. However, as we shall see, solutions of the Euler equation behave very differently from solutions of the Navier-Stokes equation, even when ν is small.

We want to understand the solutions of the Euler equations (1) and (3), or the Navier-Stokes equations (2) and (3), together with an initial condition

$$u(x) = u^0(x) \quad \text{for all } x \in \mathbb{R}^d, \quad (4)$$

where $u^0(x)$ is a given initial velocity, i.e., a vector-valued function on \mathbb{R}^d . For consistency with (3), we assume that

$$\operatorname{div} u^0(x) = 0 \quad \text{for all } x \in \mathbb{R}^d.$$

Also, to avoid physically unreasonable conditions, such as infinite energy, we demand that $u^0(x)$, as well as $u(x, t)$ for each fixed t , should tend to zero “fast enough” as $|x| \rightarrow \infty$. We will not specify here exactly what is meant by “fast enough,” but we assume from now on that we are dealing only with such rapidly decreasing velocities.

A physicist or engineer would want to know how to calculate efficiently and accurately the solution to the Navier-Stokes equations (2)–(4), and to understand how that solution behaves. A mathematician asks first whether a solution exists, and, if so, whether there is only one solution. Although the Euler equation is 250 years old and the Navier-Stokes equation well over 100 years old, there is no consensus among experts as to whether Navier-Stokes or Euler solutions exist for all time, or whether instead they “break down” at a finite time. Definitive answers supported by rigorous proofs seem a long way off.

Let us state more precisely the problem of “break-down” for the Euler and Navier-Stokes equations. Equations (1)–(3) refer to the first and second derivatives of $u(x, t)$. It is natural to suppose that the initial velocity $u^0(x)$ in (4) has derivatives

$$\partial^\alpha u^0(x) = \left(\frac{\partial}{\partial x_1} \right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_d} \right)^{\alpha_d} u^0(x)$$

of all orders, and that these derivatives tend to zero “fast enough” as $|x| \rightarrow \infty$. We then ask whether the Navier-Stokes equations (2)–(4), or the Euler equations (1), (3), and (4), have solutions $u(x, t)$, $p(x, t)$, defined for all $x \in \mathbb{R}^d$ and $t > 0$, such that the derivatives

$$\partial_{x,t}^\alpha u(x, t) = \left(\frac{\partial}{\partial t}\right)^{\alpha_0} \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_1}\right)^{\alpha_d} u(x, t)$$

and $\partial_{x,t}^\alpha p(x, t)$ of all orders exist for all $x \in \mathbb{R}^d$, $t \in [0, \infty)$ (and tend to zero “fast enough” as $|x| \rightarrow \infty$). A pair u and p with these properties is called a “smooth” solution for the Euler or Navier-Stokes equations. No one knows whether such solutions exist (in the three-dimensional case). It is known that, for some positive time $T = T(u^0) > 0$ depending on the initial velocity u^0 in (4), there exist smooth solutions $u(x, t)$, $p(x, t)$ to the Euler or Navier-Stokes equations, defined for $x \in \mathbb{R}^d$ and $t \in [0, T)$.

In two space dimensions (one speaks of “2D Euler” or “2D Navier-Stokes”), we can take $T = +\infty$; in other words, there is no “breakdown” for 2D Euler or 2D Navier-Stokes. In three space dimensions, no one can rule out the possibility that, for some finite $T = T(u^0)$ as above, there is an Euler or Navier-Stokes solution $u(x, t)$, $p(x, t)$, which is defined and smooth on

$$\Omega = \{(x, t) : x \in \mathbb{R}^3, t \in [0, T)\},$$

such that some derivative $|\partial_{x,t}^\alpha u(x, t)|$ or $|\partial_{x,t}^\alpha p(x, t)|$ is unbounded on Ω . This would imply that there is no smooth solution past time T . (We say that the 3D Navier-Stokes or Euler solution “breaks down” at time T .) Perhaps this can actually happen for 3D Euler and/or Navier-Stokes. No one knows what to believe.

Many computer simulations of the 3D Navier-Stokes and Euler equations have been carried out. Navier-Stokes simulations exhibit no evidence of breakdown, but this may mean only that initial velocities u^0 that lead to breakdown are exceedingly rare. Solutions of 3D Euler behave very wildly, so that it is hard to decide whether a given numerical study indicates a breakdown. Indeed, it is notoriously hard to perform a reliable numerical simulation of the 3D Euler equations.

It is useful to study how a Navier-Stokes or Euler solution behaves if one assumes that there is a breakdown. For instance, if there is a breakdown at time $T < \infty$ for the 3D Euler equation, then a theorem of Beale, Kato, and Majda asserts that the “vorticity”

$$\begin{aligned} \omega(x, t) &= \text{curl}(u(x, t)) \\ &= \left(\frac{\partial u_2}{\partial x_3} - \frac{\partial u_3}{\partial x_2}, \frac{\partial u_3}{\partial x_1} - \frac{\partial u_1}{\partial x_3}, \frac{\partial u_1}{\partial x_2} - \frac{\partial u_2}{\partial x_1} \right) \quad (5) \end{aligned}$$

grows so large as $t \rightarrow T$ that the integral

$$\int_0^T \left(\max_{x \in \mathbb{R}^3} |\omega(x, t)| \right) dt$$

diverges. This has been used to invalidate some plausible computer simulations that allegedly indicated a breakdown for 3D Euler. It is also known that the direction of the vorticity vector $\omega(x, t)$ must vary wildly with x , as t approaches a finite breakdown time T .

The vector ω in (5) has a natural physical meaning: it indicates how the fluid is rotating about the point x at time t . A small pinwheel placed in the fluid in position x at time t with its axis of rotation oriented parallel to $\omega(x, t)$ would be turned by the fluid at an angular velocity $|\omega(x, t)|$.

For the 3D Navier-Stokes equation, a recent result of V. Sverak shows that if there is a breakdown, then the pressure $p(x, t)$ is unbounded, both above and below.

A promising idea, pioneered by J. Leray in the 1930s, is to study “weak solutions” of the Navier-Stokes equations. The idea is as follows. At first glance, the Navier-Stokes equations (2) and (3) make sense only when $u(x, t)$, $p(x, t)$ are sufficiently smooth: for example, one would like the second derivatives of u with respect to the x_j to exist. However, a formal calculation shows that (2) and (3) are apparently equivalent to conditions that we shall call (2') and (3'), which make sense even when $u(x, t)$ and $p(x, t)$ are very rough. Let us first see how to derive (2') and (3'), and then we will discuss their use.

The starting point is the observation that a function F on \mathbb{R}^n is equal to zero if and only if $\int_{\mathbb{R}^n} F \theta \, dx = 0$ for every smooth function θ . Applying this remark to the 3D Navier-Stokes equations (2) and (3) and performing a simple formal computation (an integration by parts), we find that (2) and (3) are equivalent to the following equations:

$$\begin{aligned} &\iint_{\mathbb{R}^3 \times (0, \infty)} \left\{ - \sum_{i=1}^3 u_i \frac{\partial \theta_i}{\partial t} - \sum_{i,j=1}^3 u_i u_j \left(\frac{\partial \theta_i}{\partial x_j} \right) \right\} dx \, dt \\ &= \iint_{\mathbb{R}^3 \times (0, \infty)} \left\{ v \sum_{i,j=1}^3 \left(\frac{\partial^2}{\partial x_j^2} \theta_i \right) u_i + \left(\sum_{i=1}^3 \frac{\partial \theta_i}{\partial x_i} \right) p \right\} dx \, dt \end{aligned} \quad (2')$$

and

$$\iint_{\mathbb{R}^3 \times (0, \infty)} \left\{ \sum_{i=1}^3 u_i \frac{\partial \varphi}{\partial x_i} \right\} dx \, dt = 0. \quad (3')$$

More precisely, given any smooth functions $u(x, t)$ and $p(x, t)$, equations (2) and (3) hold if and only if (2') and (3') are satisfied for arbitrary smooth functions

$\theta_1(x, t)$, $\theta_2(x, t)$, $\theta_3(x, t)$, and $\varphi(x, t)$ that vanish outside a compact subset of $\mathbb{R}^3 \times (0, \infty)$.

We call θ_1 , θ_2 , θ_3 , and ϕ *test functions*, and we say that u and p form a *weak solution* of 3D Navier-Stokes. Since all the derivatives in (2') and (3') are applied to smooth test functions, equations (2') and (3') make sense even for very rough functions u and p . To summarize, we have the following conclusion.

A smooth pair (u, p) solves 3D Navier-Stokes if and only if it is a weak solution. However, the idea of a weak solution makes sense even for rough (u, p) .

We hope to use weak solutions, by carrying out the following plan.

Step (i): prove that suitable weak solutions exist for 3D Navier-Stokes on all of $\mathbb{R}^3 \times (0, \infty)$.

Step (ii): prove that any suitable weak solution of 3D Navier-Stokes must be smooth.

Step (iii): conclude that the suitable weak solution constructed in step (i) is in fact a smooth solution of the 3D Navier-Stokes equations on all of $\mathbb{R}^3 \times (0, \infty)$.

Here, "suitable" means "not too big"; we omit the precise definition.

Analogues of the above plan have succeeded for interesting partial differential equations. But for 3D Navier-Stokes, the plan has been only partly carried out. It has been known for a long time how to construct suitable weak solutions of 3D Navier-Stokes, but the uniqueness of these solutions has not been proved. Thanks to the work of Sheffer, of Lin, and of Caffarelli, Kohn, and Nirenberg, it is known that any suitable weak solution to 3D Navier-Stokes must be smooth (i.e., it must possess derivatives of all orders), outside a set $E \subset \mathbb{R}^3 \times (0, \infty)$ of small FRACTAL DIMENSION [III.17]. In particular, E cannot contain a curve. To rule out a breakdown, one would have to show that E is the empty set.

For the Euler equation, weak solutions again make sense, but examples due to Sheffer and Shnirelman show that they can behave very strangely. A two-dimensional fluid that is initially at rest and subject to no outside forces can suddenly start moving in a bounded region of space and then return to rest. Such behavior can occur for a weak solution of 2D Euler.

The Navier-Stokes and Euler equations give rise to a number of fundamental problems in addition to the breakdown problem discussed above. We finish this article with one such problem. Suppose that we fix an

initial velocity $u^0(x)$ for the 3D Navier-Stokes or Euler equation. The energy E_0 at time $t = 0$ is given by

$$E_0 = \frac{1}{2} \int_{\mathbb{R}^3} |u(x, 0)|^2 dx.$$

For $\nu \geq 0$, let $u^{(\nu)}(x, t) = (u_1^{(\nu)}, u_2^{(\nu)}, u_3^{(\nu)})$ denote the Navier-Stokes solution with initial velocity u^0 and with viscosity ν . (If $\nu = 0$, then $u^{(0)}$ is an Euler solution.) We assume that $u^{(\nu)}$ exists for all time, at least when $\nu > 0$. The energy for $u^{(\nu)}(x, t)$ at time $t \geq 0$ is given by

$$E^{(\nu)}(t) = \frac{1}{2} \int_{\mathbb{R}^3} |u^{(\nu)}(x, t)|^2 dx.$$

An elementary calculation based on (1)–(3) (we multiply (1) or (2) by $u_i(x)$, sum over i , integrate over all $x \in \mathbb{R}^3$, and integrate by parts) shows that

$$\frac{d}{dt} E^{(\nu)}(t) = -\frac{1}{2} \nu \int_{\mathbb{R}^3} \sum_{i,j=1}^3 \left(\frac{\partial u_i^{(\nu)}}{\partial x_j} \right)^2 dx. \quad (6)$$

In particular, for the Euler equation we have $\nu = 0$, and (6) shows that the energy is equal to E_0 , independently of time, as long as the solution exists.

Now suppose that ν is small but nonzero. From (6) it is natural to guess that $|(d/dt)E^{(\nu)}(t)|$ is small when ν is small, so that the energy remains almost constant for a long time. However, numerical and physical experiments suggest strongly that this is not the case. Instead, it seems that there exists $T_0 > 0$, depending on u^0 but independent of ν , such that the fluid loses at least half of its initial energy by time T_0 , regardless of how small ν is (provided that $\nu > 0$).

It would be very important if one could prove (or disprove) this assertion. We need to understand why a tiny viscosity dissipates a lot of energy.

III.24 Expanders

Avi Wigderson

1 The Basic Definition

An expander is a special sort of GRAPH [III.34] that has remarkable properties and many applications. Roughly speaking, it is a graph that is very hard to disconnect because every set of vertices in the graph is joined by many edges to its complement. More precisely, we say that a graph with n vertices is a c -*expander* if for every $m \leq \frac{1}{2}n$ and every set S of m vertices there are at least cm edges between S and the complement of S .

This definition is particularly interesting when G is sparse: in other words, when G has few edges. We shall concentrate on the important special case where G is

regular of degree d for some fixed constant d that is independent of the number n of vertices: this means that every vertex is joined to exactly d others. When G is regular of degree d , the number of edges from S to its complement is obviously at most dm , so if c is some fixed constant (that is, not tending to zero with n), then the number of edges between any set of vertices and its complement is within a constant of the largest number possible. As this comment suggests, we are usually interested not in single graphs but in infinite families of graphs: we say that an infinite family of d -regular graphs is a *family of expanders* if there is a constant $c > 0$ such that each graph in the family is a c -expander.

2 The Existence of Expanders

The first person to prove that expanders exist was Pinkser, who proved that if n is large and $d \geq 3$, then almost every d -regular graph with n vertices is an expander. That is, he proved that there is a constant $c > 0$ such that for every fixed $d \geq 3$, the proportion of d -regular graphs with n vertices that are *not* expanders tends to zero as n tends to infinity. This proof was an early example of the PROBABILISTIC METHOD [IV.23 §3] in combinatorics. It is not hard to see that if a d -regular graph is chosen uniformly at random, then the *expected* number of edges leaving a set S is $d|S|(n - |S|)/n$, which is at least $(\frac{1}{2}d)|S|$. Standard “tail estimates” are then used to prove that, for any fixed S , the probability that the number of edges leaving S is significantly different from its expected value is extremely small: so small that if we add up the probabilities for all sets, then even the sum is small. So with high probability all sets S have at least $c|S|$ edges to their complement. (In one respect this description is misleading: it is not a straightforward matter to discuss probabilities of events concerning random d -regular graphs because the edges are not independently chosen. However, Bollobás has defined an equivalent model for random regular graphs that allows them to be handled.)

Note that this proof does not give us an explicit description of any expander: it merely proves that they exist in abundance. This is a drawback to the proof, because, as we shall see later, there are applications for expanders that depend on some kind of explicit description, or at least on an efficient method of producing expanders. But what exactly is an “explicit description” or an “efficient method”? There are many possible answers to this question, of which we shall dis-

cuss two. The first is to demand that there is an algorithm that can list, for any integer n , all the vertices and edges of a d -regular c -expander with around n vertices (we could be flexible about this and ask for the number of vertices to be between n and n^2 , say) in a time that is polynomial in n . (See COMPUTATIONAL COMPLEXITY [IV.21 §2] for a discussion of polynomial-time algorithms.) Descriptions of this kind are sometimes called “mildly explicit.”

To get an idea of what is “mild” about this, consider the following graph. Its vertices are all 01 sequences of length k , and two such sequences are joined by an edge if they differ in exactly one place. This graph is sometimes called the *discrete cube* in k dimensions. It has 2^k vertices, so the time taken to list all the vertices and edges will be huge compared with k . However, for many purposes we do not actually need such a list: what matters is that there is a concise way of representing each vertex, and an efficient algorithm for listing the (representations of the) neighbors of any given vertex. Here the 01 sequence itself is a very concise representation, and given such a sequence σ it is very easy to list, in a time that is polynomial in k rather than 2^k , the k sequences that can be obtained by altering σ in one place. Graphs that can be efficiently described in this way (so that listing the neighbors of a vertex takes a time that is polynomial in the *logarithm* of the number of vertices) are called *strongly explicit*.

The quest for explicitly constructed expanders has been the source of some beautiful mathematics, which has often used ideas from fields such as number theory and algebra. The first explicit expander was discovered by Margulis. We give his construction and another one; we stress that although these constructions are very simple to describe, it is rather less easy to prove that they really are expanders.

Margulis's construction gives an 8-regular graph G_m for every integer m . The vertex set is $\mathbb{Z}_m \times \mathbb{Z}_m$, where \mathbb{Z}_m is the set of all integers mod m . The neighbors of the vertex (x, y) are $(x + y, y)$, $(x - y, y)$, $(x, y + x)$, $(x, y - x)$, $(x + y + 1, y)$, $(x - y + 1, y)$, $(x, y + x + 1)$, $(x, y - x + 1)$ (all operations are mod m). Margulis's proof that G_m is an expander was based on REPRESENTATION THEORY [IV.12] and did not provide any specific bound on the expansion constant c . Gabber and Galil later derived such a bound using HARMONIC ANALYSIS [IV.18]. Note that this family of graphs is strongly explicit.

Another construction provides, for each prime p , a 3-regular graph with p vertices. This time the vertex

set is \mathbb{Z}_p , and a vertex x is connected to $x + 1$, $x - 1$, and x^{-1} (where this is the inverse of $x \bmod p$, and we define the inverse of 0 to be 0). The proof that these graphs are expanders depends on a deep result in number theory, called the Selberg 3/16 theorem. This family is only mildly explicit, since we are at present unable to generate large primes deterministically.

Until recently, the only known methods for explicitly constructing expanders were algebraic. However, in 2002 Reingold, Vadhan, and Wigderson introduced the so-called zigzag product of graphs, and used it to give a combinatorial, iterative construction of expanders.

3 Expanders and Eigenvalues

The condition that a graph should be a c -expander involves all subsets of the vertices. Since there are exponentially many subsets, it would seem on the face of it that checking whether a graph is a c -expander is an exponentially long task. And, indeed, this problem turns out to be **CO-NP COMPLETE** [IV.21 §§3, 4]. However, we shall now describe a closely related property that can be checked in polynomial time, and which is in some ways more natural.

Given a graph G with n vertices, its *adjacency matrix* A is the $n \times n$ matrix where A_{uv} is defined to be 1 if u is joined to v and 0 otherwise. This matrix is real and symmetric, and therefore has n real **EIGENVALUES** [I.3 §4.3] $\lambda_1, \lambda_2, \dots, \lambda_n$, which we name in such a way that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$. Moreover, **EIGENVECTORS** [I.3 §4.3] with distinct eigenvalues are orthogonal.

It turns out that these eigenvalues encode a great deal of useful information about G . But before we come to this, let us briefly consider how A acts as a linear map. If we are given a function f , defined on the vertices of G , then Af is the function whose value at u is the sum of $f(v)$ over all neighbors v of u . From this we see immediately that if G is d -regular and f is the function that is 1 at every vertex, then Af is the function that is d at every vertex. In other words, a constant function is an eigenvector of A with eigenvalue d . It is also not hard to see that this is the largest possible eigenvalue λ_1 , and that if the graph is connected, then the second largest eigenvalue λ_2 will be strictly less than d .

In fact, the relationship between λ_2 and connectivity properties of the graph is considerably deeper than this: roughly speaking, the further away λ_2 is from d , the bigger the expansion parameter c of the graph. More precisely, it can be shown that c lies between $\frac{1}{2}(d - \lambda_2)$ and $\sqrt{2d(d - \lambda_2)}$. From this it follows that

an infinite family of d -regular graphs is a family of expanders if and only if there is some constant $a > 0$ such that the *spectral gaps* $d - \lambda_2$ are at least a for every graph in the family. One of the many reasons these bounds on c are important is that although, as we have remarked, it is hard to test whether a graph is a c -expander, its second largest eigenvalue can be computed in polynomial time. So we can at least obtain estimates for how good the expansion properties of a graph are.

Another important parameter of a d -regular graph G is the largest absolute value of any eigenvalue apart from λ_1 , which we denote by $\lambda(G)$. If $\lambda(G)$ is small, then G behaves in many respects like a random d -regular graph. For example, let A and B be two disjoint sets of vertices. If G were random, a small calculation shows that we would expect the number $E(A, B)$ of edges from A to B to be about $d|A||B|/n$. It can be shown that, for any two disjoint sets in any d -regular graph G , $E(A, B)$ will differ from this expected amount by at most $\lambda(G)\sqrt{|A||B|}$. Therefore, if $\lambda(G)$ is a small fraction of d , then between any two reasonably large sets A and B we get roughly the number of edges that we expect. This shows that graphs for which $\lambda(G)$ is small “behave like random graphs.”

It is natural to ask how small $\lambda(G)$ can be in d -regular graphs. Alon and Boppana proved that it was always at least $2\sqrt{d-1} - g(n)$ for a certain function g that tends to zero as n increases. Friedman proved that almost all d -regular graphs G with n vertices have $\lambda(G) \leq 2\sqrt{d-1} + h(n)$, where $h(n)$ tends to zero, so a typical d -regular graph comes very close to matching the best possible bound for $\lambda(G)$. The proof was a tour de force. Even more remarkably, it is possible to match the lower bound with *explicit* constructions: the famous Ramanujan graphs of Lubotzky, Philips, and Sarnak, and, independently, Margulis. They constructed, for each d such that $d-1$ is a prime power, a family of d -regular graphs G with $\lambda(G) = 2\sqrt{d-1}$.

4 Applications of Expanders

Perhaps the most obvious use for expanders is in communication networks. The fact that expanders are highly connected means that such a network is highly “fault tolerant,” in the sense that one cannot cut off part of the network without destroying a large number of individual communication lines. Further desirable properties of such a network, such as a small diameter, follow from an analysis of random walks on expanders.

T&T note: ensure that ‘NP’ is smallcaps, rather than full caps, in all cross-references before CRC.

A *random walk* of length m on a d -regular graph G is a path v_0, v_1, \dots, v_m , where each v_i is a randomly chosen neighbor of v_{i-1} . Random walks on graphs can be used to model many phenomena, and one of the questions one frequently asks about a random walk is how rapidly it “mixes.” That is, how large does m have to be before the probability that $v_m = v$ is approximately the same for all vertices v ?

If we let $p_k(v)$ be the probability that $v_k = v$, then it is not hard to show that $p_{k+1} = d^{-1}Ap_k$. In other words, the *transition matrix* T of the random walk, which tells you how the distribution after $k+1$ steps depends on the distribution after k steps, is d^{-1} times the adjacency matrix A . Therefore, its largest eigenvalue is 1, and if $\lambda(G)$ is small then all other eigenvalues are small.

Suppose that this is the case, and let p be any PROBABILITY DISTRIBUTION [III.73] on the vertices of G . Then we can write p as a linear combination $\sum_i u_i$, where u_i is an eigenvector of T with eigenvalue $d^{-1}\lambda_i$. If T is applied k times, then the new distribution will be $\sum_i (d^{-1}\lambda_i)^k u_i$. If $\lambda(G)$ is small, then $(d^{-1}\lambda_i)^k$ tends rapidly to zero, except that it equals 1 when $i = 1$. In other words, after a short time, the “nonconstant part” of p goes to zero and we are left with the uniform distribution.

Thus, random walks on expanders mix rapidly. This property is at the heart of some of the applications of expanders. For example, suppose that V is a large set, f is a function from V to the interval $[0, 1]$, and we wish to estimate quickly and accurately the average of f . A natural idea is to choose a random sample v_1, v_2, \dots, v_k of points in V and calculate the average $k^{-1} \sum_{i=1}^k f(v_i)$. If k is large and the v_i are chosen independently, then it is not too hard to prove that this sample average will almost certainly be close to the true average: the probability that they differ by more than ϵ is at most $e^{-\epsilon^2 k}$.

This idea is very simple, but actually implementing it requires a source of randomness. In theoretical computer science, randomness is regarded as a resource, and it is desirable to use less of it if one can. The above procedure needed about $\log(|V|)$ bits of randomness for each v_i , so $k \log(|V|)$ bits in all. Can we do better? Ajtai, Komlos, and Szemerédi showed that the answer is yes: big time! What one does is associate V with the vertices of an explicit expander. Then, instead of choosing v_1, v_2, \dots, v_k independently, one chooses them to be the vertices of a random walk in this expanding graph, starting at a random point v_1

of V . The randomness needed for this is far smaller: $\log(|V|)$ bits for v_1 and $\log(d)$ bits for each further v_i , making $\log(|V|) + k \log(d)$ bits in all. Since V is very large and d is a fixed constant, this is a big saving: we essentially pay only for the first sample point.

But is this sample any good? Clearly there is a heavy dependence between the v_i . However, it can be shown that *nothing* is lost in accuracy: again, the probability that the estimate differs from the true mean by more than ϵ is at most $e^{-\epsilon^2 k}$. Thus, there are no costs attached to the big saving in randomness.

This is just one of a huge number of applications of expanders, which include both practical applications and applications in pure mathematics. For instance, they were used by Gromov to give counterexamples to certain variants of the famous BAUM-CONNES CONJECTURE [IV.19 §4.4]. And certain bipartite graphs called “lossless expanders” have been used to produce linear codes with efficient decodings. (See RELIABLE TRANSMISSION OF INFORMATION [VII.6] for a description of what this means.)

PUP: I can confirm that this sentence is how it should be.

III.25 The Exponential and Logarithmic Functions

1 Exponentiation

The following is a very well-known mathematical sequence: 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, ... Each term in this sequence is twice the term before, so, for instance, 128, the seventh term in the sequence, is equal to $2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2$. Since repeated multiplications of this kind occur throughout mathematics, it is useful to have a less cumbersome notation for them, so $2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2$ is normally written as 2^7 , which we read as “2 to the power 7” or just “2 to the 7.” More generally, if a is any real number and m is any positive integer, then a^m stands for $a \times a \times \dots \times a$, where there are m a s in the product. This product is called “ a to the m ,” and numbers of the form a^m are called the *powers* of a .

The process of raising a number to a power is known as *exponentiation*. (The number m is called the *exponent*.) A fundamental fact about exponentiation is the following identity:

$$a^{m+n} = a^m \cdot a^n$$

This says that exponentiation “turns addition into multiplication.” It is easy to see why this identity must be

true if one looks at a small example and temporarily reverts to the old, cumbersome notation. For instance,

$$\begin{aligned} 2^7 &= 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \\ &= (2 \times 2 \times 2) \times (2 \times 2 \times 2 \times 2) \\ &= 2^3 \times 2^4. \end{aligned}$$

Suppose now that we are asked to evaluate $2^{3/2}$. At first sight, the question seems misconceived: an essential part of the definition of 2^m that has just been given was that m was a positive integer. The idea of multiplying one-and-a-half 2s together does not make sense. However, mathematicians like to generalize, and even if we cannot immediately make sense of 2^m except when m is a positive integer, there is nothing to stop us inventing a meaning for it for a wider class of numbers.

The more natural we make our generalization, the more interesting and useful it is likely to be. And the way we make it natural is to ensure that at all costs we keep the property of “turning addition into multiplication.” This, it turns out, leaves us with only one sensible choice for what $2^{3/2}$ should be. If the fundamental property is to be preserved, then we must have

$$2^{3/2} \cdot 2^{3/2} = 2^{3/2+3/2} = 2^3 = 8.$$

Therefore, $2^{3/2}$ has to be $\pm\sqrt{8}$. It turns out to be convenient to take $2^{3/2}$ to be positive, so we define $2^{3/2}$ to be $\sqrt{8}$.

A similar argument shows that 2^0 should be defined to be 1: if we wish to keep the fundamental property, then

$$2 = 2^1 = 2^{1+0} = 2^1 \cdot 2^0 = 2 \cdot 2^0.$$

Dividing both sides by 2 gives the answer $2^0 = 1$.

What we are doing with these kinds of arguments is solving a *functional equation*, that is, an equation where the unknown is a function. So that we can see this more clearly, let us write $f(t)$ for 2^t . The information we are given is the fundamental property $f(t+u) = f(t)f(u)$ together with one value, $f(1) = 2$, to get us started. From this we wish to deduce as much as we can about f .

It is a nice exercise to show that the two conditions we have placed on f determine the value of f at every rational number, at least if f is assumed to be positive. For instance, to show that $f(0)$ should be 1, we note that $f(0)f(1) = f(1)$, and we have already shown that $f(3/2)$ must be $\sqrt{8}$. The rest of the proof is in a similar spirit to these arguments, and the conclusion is that $f(p/q)$ must be the q th root of 2^p . More generally, the only sensible definition of $a^{p/q}$ is the q th root of a^p .

We have now extracted everything we can from the functional equation, but we have made sense of a^t only

if t is a rational number. Can we give a sensible definition when t is irrational? For example, what would be the most natural definition of $2^{\sqrt{2}}$? Since the functional equation alone does not determine what $2^{\sqrt{2}}$ should be, the way to answer a question like this is to look for some natural additional property that f might have that would, together with the functional equation, specify f uniquely. It turns out that there are two obvious choices, both of which work. The first is that f should be an *increasing* function: that is, if s is less than t , then $f(s)$ is less than $f(t)$. Alternatively, one can assume that f is CONTINUOUS [I.3 §5.2].

Let us see how the first property can in principle be used to work out $2^{\sqrt{2}}$. The idea is not to calculate it directly but to obtain better and better *estimates*. For instance, since $1.4 < \sqrt{2} < 1.5$ the order property tells us that $2^{\sqrt{2}}$ should lie between $2^{7/5}$ and $2^{3/2}$, and in general that if $p/q < \sqrt{2} < r/s$ then $2^{\sqrt{2}}$ should lie between $2^{p/q}$ and $2^{r/s}$. It can be shown that if two rational numbers p/q and r/s are very close to each other, then $2^{p/q}$ and $2^{r/s}$ are also close. It follows that as we choose fractions p/q and r/s that are closer and closer together, so the resulting numbers $2^{p/q}$ and $2^{r/s}$ converge to some limit, and this limit we call $2^{\sqrt{2}}$.

2 The Exponential Function

One of the hallmarks of a truly important concept in mathematics is that it can be defined in many different but equivalent ways. The exponential function $\exp(x)$ very definitely has this property. Perhaps the most basic way to think of it, though for most purposes not the best, is that $\exp(x) = e^x$, where e is a number whose decimal expansion begins 2.7182818. Why do we focus on this number? One property that singles it out is that if we differentiate the function $\exp(x) = e^x$, then we obtain e^x again—and e is the only number for which that is true. Indeed, this leads to a second way of defining the exponential function: it is the only solution of the differential equation $f'(x) = f(x)$ that satisfies the initial condition $f(0) = 1$.

A third way to define $\exp(x)$, and one that is often chosen in textbooks, is as the limit of a power series:

$$\exp(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots,$$

known as the *Taylor series* of $\exp(x)$. It is not immediately obvious that the right-hand side of this definition gives us some number raised to the power x , which is why we are using the notation $\exp(x)$ rather than e^x . However, with a bit of work one can verify that it

yields the basic properties $\exp(x+y) = \exp(x)\exp(y)$, $\exp(0) = 1$, and $(d/dx)\exp(x) = \exp(x)$.

There is yet another way to define the exponential function, and this one comes much closer to telling us what it really means. Suppose you wish to invest some money for ten years and are given the following choice: either you can add 100% to your investment (that is, double it) at the end of the ten years, or each year you can take whatever you have and increase it by 10%. Which would you prefer?

The second is the better investment because in the second case the interest is *compounded*: for instance, if you start with \$100, then after a year you will have \$110 and after two years you will have \$121. The increase of \$11 in the second year breaks down as 10% interest on the original \$100 plus a further dollar, which is 10% interest on the interest earned in the first year. Under the second scheme, the amount of money you end up with is \$100 times $(1.1)^{10}$, since each year it multiplies by 1.1. The approximate value of $(1.1)^{10}$ is 2.5937, so you will get almost \$260 instead of \$200.

What if you compounded your interest monthly? Instead of multiplying your investment by $1\frac{1}{10}$ ten times, you would multiply it by $1\frac{1}{120}$ 120 times. By the end of ten years your \$100 would have been multiplied by $(1 + \frac{1}{120})^{120}$, which is approximately 2.707. If you compounded it daily, you could increase this to approximately 2.718, which is suspiciously close to e . In fact, e can be defined as the limit, as n tends to infinity, of the number $(1 + \frac{1}{n})^n$.

It is not instantly obvious that this expression really does tend to a limit. For any fixed power m , the limit of $(1 + \frac{1}{n})^m$ as n tends to infinity is 1, while for any fixed n , the limit as m tends to infinity is ∞ . When it comes to $(1 + \frac{1}{n})^n$, the increase in the power just compensates for the decrease in the number $1 + \frac{1}{n}$ and we get a limit between 2 and 3. If x is any real number, then $(1 + \frac{x}{n})^n$ also converges to a limit, and this we define to be $\exp(x)$.

Here is a sketch of an argument that shows that if we define $\exp(x)$ this way, then $\exp(x)\exp(y) = \exp(x+y)$, the main property we need if our definition is to be a good one. Let us take a very large n and look at the number

$$\left(1 + \frac{x}{n}\right)^n \left(1 + \frac{y}{n}\right)^n,$$

which equals

$$\left(1 + \frac{x}{n} + \frac{y}{n} + \frac{xy}{n^2}\right)^n.$$

Now the ratio of $1 + x/n + y/n + xy/n^2$ to $1 + x/n + y/n$ is smaller than $1 + xy/n^2$, and $(1 + xy/n^2)^n$ can be shown to converge to 1 (as here the increase in n is not enough to compensate for the rapid decrease in xy/n^2). Therefore, for large n the number we have is very close to

$$\left(1 + \frac{x+y}{n}\right)^n.$$

Letting n tend to infinity, we deduce the result.

3 Extending the Definition to Complex Numbers

If we think of $\exp(x)$ as e^x , then the idea of generalizing the definition to complex numbers seems hopeless: our intuition tells us nothing, the functional equation does not help, and we cannot use continuity or order relations to determine it for us. However, both the power series and the compound-interest definitions can be generalized easily. If z is a complex number, then the most usual definition of $\exp(z)$ is

$$1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots.$$

Setting $z = i\theta$, for a real number θ , and splitting the resulting expression into its real and imaginary parts, we obtain

$$1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} + \cdots + i\left(\theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \cdots\right),$$

which, using the power-series expansions for $\cos(\theta)$ and $\sin(\theta)$, tells us that $\exp(i\theta) = \cos(\theta) + i\sin(\theta)$, the formula for the point with argument θ on the unit circle in the complex plane. In particular, if we take $\theta = \pi$, we obtain the famous formula $e^{i\pi} = -1$ (since $\cos(\pi) = -1$ and $\sin(\pi) = 0$).

This formula is so striking that one feels that it ought to hold for a good reason, rather than being a mere fact that one notices after carrying out some formal algebraic manipulations. And indeed there is a good reason. To see it, let us return to the compound-interest idea and define $\exp(z)$ to be the limit of $(1 + z/n)^n$ as n tends to infinity. Let us concentrate just on the case where $z = i\pi$: why should $(1 + i\pi/n)^n$ be close to -1 when n is very large?

To answer this, let us think geometrically. What is the effect on a complex number of multiplying it by $1 + i\pi/n$? On the Argand diagram this number is very close to 1 and vertically above it. Because the vertical line through 1 is tangent to the circle, this means that the number is very close indeed to a number that lies on the circle and has argument π/n (since the argument

of a number on the circle is the length of the circular arc from 1 to that number, and in this case the circular arc is almost straight). Therefore, multiplication by $1 + i\pi/n$ is very well approximated by rotation through an angle of π/n . Doing this n times results in a rotation by π , which is the same as multiplication by -1 . The same argument can be used to justify the formula $\exp(i\theta) = \cos(\theta) + i\sin(\theta)$.

Continuing in this vein, let us see why the derivative of the exponential function is the exponential function. We know already that $\exp(z + w) = \exp(z)\exp(w)$, so the derivative of \exp at z is the limit as w tends to zero of $\exp(z)(\exp(w) - 1)/w$. It is therefore enough to show that $\exp(w) - 1$ is very close to w when w is small. To get a good idea of $\exp(w)$ we should take a large n and consider $(1 + w/n)^n$. It is not hard to prove that this is indeed close to $1 + w$, but here is an informal argument instead. Suppose that you have a bank account that offers a tiny rate of interest over a year, say 0.5%. How much better would you do if you could compound this interest monthly? The answer is not very much: if the total amount of interest is very small, then the interest on the interest is negligible. This, in essence, is why $(1 + w/n)^n$ is approximately $1 + w$ when w is small.

One can extend the definition of the exponential function yet further. The main ingredients one needs are addition, multiplication, and the possibility of limiting arguments. So, for example, if x is an element of a BANACH ALGEBRA [III.12] A , then $\exp(x)$ makes sense. (Here, the power series definition is the easiest, though not necessarily the most enlightening.)

4 The Logarithm Function

Natural logarithms, like exponentials, can be defined in many ways. Here are three.

- (i) The function \log is the inverse of the function \exp . That is, if t is a positive real number, then the statement $u = \log(t)$ is equivalent to the statement $t = \exp(u)$.
- (ii) Let t be a positive real number. Then

$$\log(t) = \int_1^t \frac{dx}{x}.$$

- (iii) If $|x| < 1$ then $\log(1 + x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \dots$. This defines $\log(t)$ for $0 < t < 2$. If $t \geq 2$ then $\log(t)$ can be defined as $-\log(1/t)$.

The most important feature of the logarithmic function is a functional equation that is the reverse of the

functional equation for \exp , namely $\log(st) = \log(s) + \log(t)$. That is, whereas \exp turns addition into multiplication, \log turns multiplication into addition. A more formal way of putting this is that \mathbb{R} forms a group under addition, and \mathbb{R}_+ , the set of positive real numbers, forms a group under multiplication. The function \exp is an isomorphism from \mathbb{R} to \mathbb{R}_+ , and \log , its inverse, is an isomorphism from \mathbb{R}_+ to \mathbb{R} . Thus, in a sense the two groups have the same structure, and the exponential and logarithmic functions demonstrate this.

Let us use the first definition of \log to see why $\log(st)$ must equal $\log(s) + \log(t)$. Write $s = \exp(a)$ and $t = \exp(b)$. Note that $a = \log(s)$ and $b = \log(t)$. Then $\log(s) = a$, $\log(t) = b$, and

$$\begin{aligned}\log(st) &= \log(\exp(a)\exp(b)) \\ &= \log(\exp(a + b)) \\ &= a + b.\end{aligned}$$

The result follows.

In general, the properties of \log closely follow those of \exp . However, there is one very important difference, which is a complication that arises when one tries to extend \log to the complex numbers. At first it seems quite easy: every complex number z can be written as $re^{i\theta}$ for some nonnegative real number r and some θ (the modulus and argument of z , respectively). If $z = re^{i\theta}$ then $\log(z)$, one might think, should be $\log(r) + i\theta$ (using the functional equation for \log and the fact that \log inverts \exp). The problem with this is that θ is not uniquely determined. For instance, what is $\log(1)$? Normally we would like to say 0, but we could, perversely, say that $1 = e^{2\pi i}$ and claim that $\log(1) = 2\pi i$.

Because of this difficulty, there is no single best way to define the logarithmic function on the entire complex plane, even if 0, a number that does not have a logarithm however you look at it, is removed. One convention is to write $z = re^{i\theta}$ with $r > 0$ and $0 \leq \theta < 2\pi$, which can be done in exactly one way, and then define $\log(z)$ to be $\log(r) + i\theta$. However, this function is not continuous: as you cross the positive real axis, the argument jumps by 2π and the logarithm jumps by $2\pi i$.

Remarkably, this difficulty, far from being a blow to mathematics, is an entirely positive phenomenon that lies behind several remarkable theorems in complex analysis, such as Cauchy's residue theorem, which allows one to evaluate very general path integrals.

III.26 The Fast Fourier Transform

If f is a periodic function with period 1, then one can obtain a great deal of useful information about f by calculating its Fourier coefficients (see THE FOURIER TRANSFORM [III.27] for a discussion of why). This is true for both theoretical and practical reasons, and because of the latter it is highly desirable to have a good way of computing Fourier coefficients quickly.

The r th Fourier coefficient of f is given by the formula

$$\hat{f}(r) = \int_0^1 f(x) e^{-2\pi i r x} dx.$$

If we do not have an explicit formula for the integral (as would be the case, for instance, if f were derived from some physical signal rather than a mathematical formula), then we will want to approximate this integral numerically, and a natural way to do that is to *discretize* it: that is, turn it into a sum of the form $N^{-1} \sum_{n=0}^{N-1} f(n/N) e^{-2\pi i r n/N}$. If f is not too wildly oscillating and r is not too big, then this should be a good approximation.

The sum above will be unchanged if we add a multiple of N to r , so we now care only about the values of f at points of the form n/N . Moreover, the periodicity of f tells us that adding a multiple of N to n also makes no difference. So we can regard both n and r as belonging to the group \mathbb{Z}_N of integers mod N (see MODULAR ARITHMETIC [III.60]). Let us change our notation to one that reflects this. Given a function g defined on \mathbb{Z}_N we define the *discrete Fourier transform* of g to be the function \hat{g} , also defined on \mathbb{Z}_N , which is given by the formula

$$\hat{g}(r) = N^{-1} \sum_{n \in \mathbb{Z}_N} g(n) \omega^{-rn}, \quad (1)$$

where we are writing ω for $e^{2\pi i/N}$, so that $\omega^{-rn} = e^{-2\pi i r n/N}$. Note that the sum over n could be regarded as a sum from 0 to $N-1$ just as above; the other notational change is that we have written $g(n)$ instead of $f(n/N)$.

The discrete Fourier transform can be thought of as multiplying a column vector (corresponding to the function g) by an $N \times N$ matrix (with entries $N^{-1} \omega^{-rn}$ for each r and n). Therefore it can be calculated using about N^2 arithmetical operations. The fast Fourier transform arises from the observation that the sum in (1) has symmetry properties that allow it to be calculated much more efficiently. This is most easily seen when N is a power of 2, and to make it even easier we

shall look at the case $N = 8$. The sums to be evaluated are then

$$g(0) + \omega^r g(1) + \omega^{2r} g(2) + \cdots + \omega^{7r} g(7)$$

for each r between 0 and 7. Now a sum like this can be rewritten as

$$g(0) + \omega^{2r} g(2) + \omega^{4r} g(4) + \omega^{6r} g(6) \\ + \omega^r (g(1) + \omega^{2r} g(3) + \omega^{4r} g(5) + \omega^{6r} g(7)),$$

which is interesting because

$$g(0) + \omega^{2r} g(2) + \omega^{4r} g(4) + \omega^{6r} g(6)$$

and

$$g(1) + \omega^{2r} g(3) + \omega^{4r} g(5) + \omega^{6r} g(7)$$

are themselves values of discrete Fourier transforms. For instance, if we set $h(n) = g(2n)$ for $0 \leq n \leq 3$, and write ψ for $\omega^2 = e^{2\pi i/4}$, then the first expression equals $h(0) + \psi^r h(1) + \psi^{2r} h(2) + \psi^{3r} h(3)$. If we think of h as being defined on \mathbb{Z}_4 , then this is precisely the formula for $\hat{h}(r)$.

A similar remark applies to the second expression, so if we can calculate the discrete Fourier transforms of the “even part” of g and the “odd part” of g , then it will be very straightforward to obtain each value of the Fourier transform of g itself: it will be a linear combination of values of the transforms of the two parts of g . Thus, if N is even and we write $F(N)$ for the number of operations needed to calculate the discrete Fourier transform of a function defined on \mathbb{Z}_N , we obtain a recurrence of the form

$$F(N) = 2F(N/2) + CN.$$

The interpretation of this is that in order to work out the N values of the transform of a function on \mathbb{Z}_N , it is enough to work out two such transforms for functions on $\mathbb{Z}_{N/2}$ and work out N linear combinations.

If N is a power of 2, then we can iterate this: $F(N/2)$ will be at most $2F(N/4) + CN/2$, and so on. It is not hard to show as a result that $F(N)$ is at most $CN \log N$ for some constant C , a considerable improvement on CN^2 . If N is not a power of 2, then the above argument does not work, but there are modifications of the method that do, and that lead to similar efficiency gains. (Indeed, this is true for the Fourier transform on an arbitrary finite Abelian group.)

Once we can calculate Fourier transforms efficiently, there are other calculations that immediately become easy as well. A simple example is the *inverse* Fourier transform, which has a formula very similar to that of the Fourier transform and can therefore be calculated in a similar way. Another calculation that becomes

easy is the *convolution* of two sequences, which is defined as follows. If $a = (a_0, a_1, a_2, \dots, a_m)$ and $b = (b_0, b_1, b_2, \dots, b_n)$ are two sequences, then their convolution is the sequence $c = (c_0, c_1, c_2, \dots, c_{m+n})$, where each c_r is defined to be $a_0 b_r + a_1 b_{r-1} + \dots + a_r b_0$. This sequence is denoted by $a * b$. One of the most important properties of Fourier transforms is that they “convert convolutions into multiplication.” That is, if we find a suitable way of regarding a and b as functions on \mathbb{Z}_N , then the Fourier transform of $a * b$ is the function $r \mapsto \hat{a}(r)\hat{b}(r)$. Therefore, to work out $a * b$ we can work out \hat{a} and \hat{b} , multiply them together for each r , and take the inverse Fourier transform of the result. All stages of this calculation are quick, so calculating convolutions is quick.

This immediately leads to a quick way of multiplying the two polynomials $a_0 + a_1 x + \dots + a_m x^m$ and $b_0 + b_1 x + \dots + b_n x^n$ together, since the coefficients of the product are given by the sequence $c = a * b$. If all the a_i are between 0 and 9, it is a quick process to evaluate the product polynomial at $x = 10$ (since none of the coefficients c_r will have many digits), so we also have a method of multiplying two n -digit integers together that is far faster than long multiplication. These are two of the huge number of applications of the fast Fourier transform. A more direct source of applications occurs in engineering, where one frequently wishes to analyze a signal by looking at its Fourier transform. A very surprising application is to QUANTUM COMPUTATION [III.76]: a famous result of Peter Shor is that one can use a quantum computer to factorize large integers very quickly; this algorithm depends in an essential way on the fast Fourier transform, but uses the power of quantum computing in an almost miraculous way to divide the $N \log N$ steps into N lots of $\log N$ steps that can be carried out “in parallel.”

III.27 The Fourier Transform

Terence Tao

Let f be a function from \mathbb{R} to \mathbb{R} . Typically, there is not much that one can say about f , but certain functions have useful symmetry properties. For instance, f is called *even* if $f(-x) = f(x)$ for every x , and it is called *odd* if $f(-x) = -f(x)$ for every x . Furthermore, every function f can be written as a *superposition* of an even part, f_e , and an odd part, f_o . For instance, the function $f(x) = x^3 + 3x^2 + 3x + 1$ is neither even nor odd, but it can be written as $f_e(x) + f_o(x)$, where

$f_e(x) = 3x^2 + 1$ and $f_o(x) = x^3 + 3x$. For a general function f , the decomposition is unique and is given by the formulas $f_e(x) = \frac{1}{2}(f(x) + f(-x))$ and $f_o(x) = \frac{1}{2}(f(x) - f(-x))$.

What are the symmetry properties enjoyed by even and odd functions? A useful way to regard them is as follows. We have a group of two transformations of the real line: one is the identity map $\iota : x \mapsto x$ and the other is the reflection $\rho : x \mapsto -x$. Now any transformation ϕ of the real line gives rise to a transformation of the functions defined on the real line: given a function f , the transformed function is the function $g(x) = f(\phi(x))$. In the case at hand, if $\phi = \iota$ then the transformed function is just $f(x)$, while if $\phi = \rho$ then it is $f(-x)$. If f is either even or odd, then both the transformed functions are *scalar multiples* of the original function f . In particular, when $\phi = \rho$, the transformed function is $f(x)$ when f is even (so the scalar multiple is 1) and $-f(x)$ when f is odd (so the scalar multiple is -1).

The procedure just described can be thought of as a very simple prototype of the general notion of a Fourier transform. Very broadly speaking, a Fourier transform is a systematic way to decompose “generic” functions into a superposition of “symmetric” functions. These symmetric functions are usually quite explicitly defined: for instance, one of the most important examples is a decomposition into the TRIGONOMETRIC FUNCTIONS [III.94] $\sin(nx)$ and $\cos(nx)$. They are also often related to physical concepts such as frequency or energy. The symmetry will usually be associated with a GROUP [I.3 §2.1] G , which is usually Abelian. (In the case considered above, it is the two-element group.) Indeed, the Fourier transform is a fundamental tool in the study of groups, and more precisely in the REPRESENTATION THEORY [IV.12] of groups, which concerns different ways in which a group can be regarded as a group of symmetries. It is also related to topics in linear algebra, such as the representation of a vector as linear combinations of an ORTHONORMAL BASIS [III.37], or as linear combinations of EIGENVECTORS [I.3 §4.3] of a matrix or LINEAR OPERATOR [III.52].

For a more complicated example, let us fix a positive integer n and let us define a systematic way of decomposing functions from \mathbb{C} to \mathbb{C} , that is, complex-valued functions defined on the complex plane. If f is such a function and j is an integer between 0 and $n - 1$, then we say that f is a *harmonic of order j* if it has the following property. Let $\omega = e^{2\pi i/n}$, so that ω is a primitive n th root of 1 (meaning that $\omega^n = 1$ but no smaller

positive power of ω gives 1). Then $f(\omega z) = \omega^j f(z)$ for every $z \in \mathbb{C}$. Notice that if $n = 2$, then $\omega = -1$, so when $j = 0$ we recover the definition of an even function and when $j = 1$ we recover the definition of an odd function. In fact, inspired by this, we can give a general formula for a decomposition of f into harmonics, which again turns out to be unique. If we define

$$f_j(z) = \frac{1}{n} \sum_{k=0}^{n-1} f(\omega^k z) \omega^{-jk},$$

then it is a simple exercise to prove that

$$f(z) = \sum_{j=0}^{n-1} f_j(z)$$

for every z (use the fact that $\sum_j \omega^{-jk} = n$ if $k = 0$ and 0 otherwise), and that $f_j(\omega z) = \omega^j f_j(z)$ for every z . Thus, f can be decomposed as a sum of harmonics. The group associated with this Fourier transform is the multiplicative group of the n th roots of unity $1, \omega, \dots, \omega^{n-1}$, or the cyclic group of order n . The root ω^j is associated with the rotation of the complex plane through an angle of $2\pi j/n$.

Now let us consider infinite groups. Let f be a complex-valued function defined on the unit circle $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$. To avoid technical issues we shall assume that f is *smooth*—that is, it is infinitely differentiable. Now if f is a function of the simple form $f(z) = cz^n$ for some integer n and some constant c , then f will have rotational symmetry of order n . That is, if $\omega = e^{2\pi i/n}$ again, then $f(\omega z) = f(z)$ for all complex numbers z . After our earlier examples, it should come as no surprise that an arbitrary smooth function f can be expressed as a superposition of such rotationally symmetric functions. Indeed, one can write

$$f(z) = \sum_{n=-\infty}^{\infty} \hat{f}(n) z^n,$$

where the numbers $\hat{f}(n)$, called the *Fourier coefficients* of f at the *frequencies* n , are given by the formula

$$\hat{f}(n) = \frac{1}{2\pi} \int_0^{2\pi} f(e^{i\theta}) e^{-in\theta} d\theta.$$

This formula can be thought of as the limiting case $n \rightarrow \infty$ of the previous decomposition, restricted to the unit circle. It can also be regarded as a generalization of the Taylor series expansion of a HOLOMORPHIC FUNCTION [I.3 §5.6]. If f is holomorphic on the closed unit disk $\{z \in \mathbb{C} : |z| \leq 1\}$, then one can write

$$f(z) = \sum_{n=0}^{\infty} a_n z^n,$$

where the *Taylor coefficient* a_n is given by the formula

$$a_n = \frac{1}{2\pi i} \int_{|z|=1} \frac{f(z)}{z^{n+1}} dz.$$

In general, there are very strong links between Fourier analysis and complex analysis.

When f is smooth, then its Fourier coefficients decay to zero very quickly and it is easy to show that the Fourier series $\sum_{n=-\infty}^{\infty} \hat{f}(n) z^n$ converges. The issue becomes more subtle if f is not smooth (for instance, if it is merely continuous). Then one must be careful to specify the precise sense in which the series converges. In fact, a significant portion of HARMONIC ANALYSIS [IV.18] is devoted to questions of this kind, and to developing tools for answering them.

The group of symmetries associated with this version of Fourier analysis is the circle group \mathbb{T} . (Notice that we can think of the number $e^{i\theta}$ both as a point in the circle and as a rotation through an angle of θ . Thus, the circle can be identified with its own group of rotational symmetries.) But there is a second group that is important here as well, namely the additive group \mathbb{Z} of all integers. If we take two of our basic symmetric functions, z^m and z^n , and multiply them together, then we obtain the function z^{m+n} , so the map $n \rightarrow z^n$ is an isomorphism from \mathbb{Z} to the set of all these functions under multiplication. The group \mathbb{Z} is known as the *Pontryagin dual* to \mathbb{T} .

In the theory of partial differential equations and in related areas of harmonic analysis, the most important Fourier transform is defined on the Euclidean space \mathbb{R}^d . Among all functions $f : \mathbb{R}^d \rightarrow \mathbb{C}$, the ones considered to be “basic” are the *plane waves* $f(x) = c_\xi e^{2\pi i x \cdot \xi}$, where $\xi \in \mathbb{R}^d$ is a vector (known as the *frequency* of the plane wave), $x \cdot \xi$ is the dot product between the position x and the frequency ξ , and c_ξ is a complex number (whose magnitude is the *amplitude* of the plane wave). Notice that sets of the form $H_\lambda = \{x : x \cdot \xi = \lambda\}$ are (hyper)planes orthogonal to ξ , and on each such set the value of $f(x)$ is constant. Moreover, the value taken by f on H_λ is always equal to the value taken on $H_{\lambda+2\pi}$. This explains the name “plane waves.” It turns out that if a function f is sufficiently “nice” (e.g., smooth and rapidly decreasing as x gets large), then it can be represented uniquely as the superposition of plane waves, where a “superposition” is now interpreted as an integral rather than a summation. More precisely, we have

the formulas¹

$$f(x) = \int_{\mathbb{R}^d} \hat{f}(\xi) e^{2\pi i x \cdot \xi} d\xi,$$

where

$$\hat{f}(\xi) = \int_{\mathbb{R}^d} f(x) e^{-2\pi i x \cdot \xi} dx.$$

The function $\hat{f}(\xi)$ is known as the *Fourier transform* of f , and the second formula is known as the *Fourier inversion formula*. These two formulas show how to determine the Fourier-transformed function from the original function and vice versa. One can view the quantity $\hat{f}(\xi)$ as the extent to which the function f contains a component that oscillates at frequency ξ . As it turns out, there is no difficulty in justifying the convergence of these integrals when f is sufficiently nice, though the issue again becomes more subtle for functions that are somewhat rough or slowly decaying. In this case, the underlying group is the Euclidean group \mathbb{R}^d (which can also be thought of as the group of d -dimensional translations); note that both the position variable x and the frequency variable ξ are contained in \mathbb{R}^d , so \mathbb{R}^d is also the Pontryagin dual group in this setting.²

One major application of the Fourier transform lies in understanding various linear operations on functions, such as, for instance, the Laplacian on \mathbb{R}^d . Given a function $f : \mathbb{R}^d \rightarrow \mathbb{C}$, its Laplacian Δf is defined by the formula

$$\Delta f(x) = \sum_{j=1}^d \frac{\partial^2 f}{\partial x_j^2},$$

where we think of the vector x in coordinate form, $x = (x_1, \dots, x_d)$, and of f as a function $f(x_1, \dots, x_d)$ of d real variables. To avoid technicalities let us consider only those functions that are smooth enough for the above formula to make sense without any difficulty.

In general, there is no obvious relationship between a function f and its Laplacian Δf . But when f is a plane wave such as $f(x) = e^{2\pi i x \cdot \xi}$, then there is a very simple relationship:

$$\Delta e^{2\pi i x \cdot \xi} = -4\pi^2 |\xi|^2 e^{2\pi i x \cdot \xi}.$$

That is, the effect of the Laplacian on the plane wave $e^{2\pi i x \cdot \xi}$ is to multiply it by the scalar $-4\pi^2 |\xi|^2$. In

other words, the plane wave is an eigenfunction³ for the Laplacian Δ , with eigenvalue $-4\pi^2 |\xi|^2$. (More generally, plane waves will be eigenfunctions for any linear operation that commutes with translations.) Therefore, the Laplacian, when viewed through the lens of the Fourier transform, is very simple: the Fourier transform lets one write an arbitrary function as a superposition of plane waves, and the Laplacian has a very simple effect on each plane wave. To be explicit about it,

$$\begin{aligned} \Delta f(x) &= \Delta \int_{\mathbb{R}^d} \hat{f}(\xi) e^{2\pi i x \cdot \xi} d\xi \\ &= \int_{\mathbb{R}^d} \hat{f}(\xi) \Delta e^{2\pi i x \cdot \xi} d\xi \\ &= \int_{\mathbb{R}^d} (-4\pi^2 |\xi|^2) \hat{f}(\xi) e^{2\pi i x \cdot \xi} d\xi, \end{aligned}$$

which gives us a formula for the Laplacian of a general function. Here we have interchanged the Laplacian Δ with an integral; this can be rigorously justified for suitably nice f , but we omit the details.

This formula represents Δf as a superposition of plane waves. But any such representation is unique, and the Fourier inversion formula tells us that

$$\Delta f(x) = \int_{\mathbb{R}^d} \widehat{\Delta f}(\xi) e^{2\pi i x \cdot \xi} d\xi.$$

Therefore,

$$\widehat{\Delta f}(\xi) = (-4\pi^2 |\xi|^2) \hat{f}(\xi),$$

a fact that can also be derived directly from the definition of the Fourier transform using integration by parts. This identity shows that the Fourier transform *diagonalizes* the Laplacian: the operation of taking the Laplacian, when viewed using the Fourier transform, is nothing more than multiplication of a function $F(\xi)$ by the *multiplier* $-4\pi^2 |\xi|^2$. The quantity $-4\pi^2 |\xi|^2$ can be interpreted as the *energy level* associated⁴ with the frequency ξ . In other words, the Laplacian can be viewed as a *Fourier multiplier*, meaning that to calculate the Laplacian you take the Fourier transform, multiply by the multiplier, and then take the inverse Fourier transform again. This viewpoint allows one to manipulate the Laplacian very easily. For instance, we can iterate the above formula to compute higher powers of the Laplacian:

$$\widehat{\Delta^n f}(\xi) = (-4\pi^2 |\xi|^2)^n \hat{f}(\xi) \quad \text{for } n = 0, 1, 2, \dots$$

Indeed, we are now in a position to develop more general functions of the Laplacian. For instance, we can

1. In some texts, the Fourier transform is defined slightly differently, with factors such as 2π and -1 being moved to other places. These notational differences have some minor benefits and drawbacks, but they are all equivalent to each other.

2. This is because of our reliance on the dot product; if one did not want to use this dot product, the Pontryagin dual would instead be $(\mathbb{R}^d)^*$, the dual vector space to \mathbb{R}^d . But this subtlety is not too important in most applications.

3. Strictly speaking, this is a *generalized* eigenfunction, as plane waves are not square-integrable on \mathbb{R}^d .

4. When taking this view, it is customary to replace Δ by $-\Delta$ in order to make the energies positive.

take a square root as follows:

$$\widehat{\sqrt{-\Delta}f}(\xi) = 2\pi|\xi|\hat{f}(\xi).$$

This leads to the theory of fractional differential operators (which are in turn a special case of *pseudodifferential operators*), as well as the more general theory of FUNCTIONAL CALCULUS [IV.19 §3.1], in which one starts with a given operator (such as the Laplacian) and then studies various functions of that operator, such as square roots, exponentials, inverses, and so forth.

As the above discussion shows, the Fourier transform can be used to develop a number of interesting operations, which have particular importance in the theory of differential equations. To analyze these operations effectively, one needs various *estimates* on the Fourier transform. For instance, it is often important to know how the size of a function f , as measured by some norm, relates to the size of its Fourier transform, as measured by a possibly different norm. For a further discussion of this point, see FUNCTION SPACES [III.29]. One particularly important and striking estimate of this type is the *Plancherel identity*,

$$\int_{\mathbb{R}^d} |f(x)|^2 dx = \int_{\mathbb{R}^d} |\hat{f}(\xi)|^2 d\xi,$$

which shows that the L_2 -norm of a Fourier transform is actually *equal* to the L_2 -norm of the original function. The Fourier transform is therefore a unitary operation, so one can view the frequency-space representation of a function as being in some sense a “rotation” of the physical-space representation.

Developing further estimates related to the Fourier transform and associated operators is a major component of harmonic analysis. A variant of the Plancherel identity is the *convolution formula*:

$$\int_{\mathbb{R}^d} f(y)g(x-y) dy = \int_{\mathbb{R}^d} \hat{f}(\xi)\hat{g}(\xi)e^{2\pi i x \cdot \xi} d\xi.$$

This formula allows one to analyze the convolution $f * g(x) = \int_{\mathbb{R}^d} f(y)g(x-y) dy$ of two functions f, g in terms of their Fourier transform; in particular, if the Fourier coefficients of f or g are small, then we expect the convolution $f * g$ to be small as well. This relationship means that the Fourier transform controls certain *correlations* of a function with itself and with other functions, which makes the Fourier transform an important tool in understanding the randomness and uniform distribution properties of various objects in probability theory, harmonic analysis, and number theory. For instance, one can pursue the above ideas to establish the central limit theorem, which asserts that the sum of many independent random variables

will eventually resemble a Gaussian distribution (see PROBABILITY DISTRIBUTIONS [III.73 §5]); one can even use such methods to establish VINOGRADOV’S THEOREM [V.29], that every sufficiently large odd number is the sum of three primes.

There are many directions in which to generalize the above set of ideas. For instance, one can replace the Laplacian by a more general operator and the plane waves by (generalized) eigenfunctions of that operator. This leads to the subject of SPECTRAL THEORY [III.88] and functional calculus; one can also study the algebra of Fourier multipliers (and of convolution) more abstractly, which leads to the theory of C^* -ALGEBRAS [IV.19 §3]. One can also go beyond the theory of linear operators and study bilinear, multilinear, or even fully nonlinear operators. This leads in particular to the theory of *paraproducts*, which are generalizations of the pointwise product operation $(f(x), g(x)) \mapsto fg(x)$ that are of importance in differential equations. In another direction, one can replace Euclidean space \mathbb{R}^d by a more general group, in which case the notion of a plane wave is replaced by the notion of a *character* (if the group is Abelian) or a *representation* (if the group is non-Abelian). There are other variants of the Fourier transform, such as the Laplace transform or the Mellin transform (for more about other transforms, see the article TRANSFORMS [III.93]), which are very similar algebraically to the Fourier transform and play similar roles (for instance, the Laplace transform is also useful in analyzing differential equations). We have already seen that Fourier transforms are connected to Taylor series; there is also a connection to some other important series expansions, notably Dirichlet series, as well as expansions of functions in terms of SPECIAL POLYNOMIALS [III.87] such as orthogonal polynomials or SPHERICAL HARMONICS [III.89].

The Fourier transform decomposes a function exactly into many components, each of which has a precise frequency. In some applications it is more useful to adopt a “fuzzier” approach, in which a function is decomposed into fewer components but each component has a range of frequencies rather than consisting purely of a single frequency. Such decompositions can have the advantage of being less constrained by the *uncertainty principle*, which asserts that it is impossible for both a function and its Fourier transform to be concentrated in very small regions of \mathbb{R}^d . This leads to some variants of the Fourier transform, such as WAVELET TRANSFORMS [VII.3], which are better suited to a number of problems in applied and

computational mathematics, and also to certain questions in harmonic analysis and differential equations. The uncertainty principle, being fundamental to quantum mechanics, also connects the Fourier transform to mathematical physics, and in particular to the connections between classical and quantum physics, which can be studied rigorously using the methods of geometric quantization and microlocal analysis.

III.28 Fuchsian Groups

Jeremy Gray

One of the most basic objects in geometry is the *torus*: a surface that has the shape of the surface of a bagel. If you want to construct one, you can do so by taking a square and gluing opposite edges together. When you glue the top and bottom edges together you have a cylinder, and when you glue the other two edges together, which have now become circles, you obtain your torus.

A more mathematical way of making a torus is as follows. We start with the usual (x, y) coordinate plane and the square in it with vertices at $(0, 0)$, $(1, 0)$, $(1, 1)$, and $(0, 1)$, which consists of the points whose coordinates satisfy $0 \leq x \leq 1$, $0 \leq y \leq 1$. This square can be moved around horizontally and vertically. If we shift it m units horizontally and n units vertically, where m and n are integers, we get the square that consists of the points whose coordinates satisfy $m \leq x \leq m + 1$, $n \leq y \leq n + 1$. As m and n run through all the integers, we see that the copies of the square cover the whole plane, with four squares coming together at each point with integer coordinates. The plane is said to be *tiled* or *tessellated* (from the Latin word for a marble chip in a mosaic), and it is easy to see that you can color the squares alternately black and white and get an infinite checkerboard pattern.

To make the torus we “identify” points. We say that the points (x, y) and (x', y') correspond to the same point in a certain new figure if $x - x'$ and $y - y'$ are both integers. To see what the new figure looks like, we observe that any point in the plane corresponds to a point inside, or on the edge of, our original square. Moreover, the point (x, y) corresponds to exactly one point inside the square provided that neither x nor y is an integer. So our new space looks a lot like our original square. But what about the points $(\frac{1}{4}, 0)$ and $(\frac{1}{4}, 1)$? They correspond to the same point in our new space, as do any corresponding pairs of points on the upper and lower edges of our square. So those edges are identified

in our new space. By a similar argument, so too are the left and right edges. The result is that, after points are identified according to our rule, we obtain the torus.

If we make the torus in this way, we can draw small figures on it just by drawing them in the original square; lengths in the square will then correspond exactly to lengths on the torus. This is how old-fashioned printing on a drum works: an inked figure on a cylinder is rolled over the paper to make exact copies of the figure. Thus, as far as small figures are concerned, the geometry of the torus is exactly like Euclidean geometry. In mathematical language we say that the geometry on the torus is induced from the geometry on the plane, and therefore that it is *locally Euclidean*. Globally, of course, it is different, because one can draw curves on the torus that cannot be shrunk to a point, whereas one cannot do so on the plane.

Notice, too, that we have brought in a group to do the bulk of the work for us. In this case the group is the set of all pairs (m, n) where m and n are integers, with $(m, n) + (m', n')$ defined to be $(m + m', n + n')$.

The torus and the sphere are but two of an infinite class of surfaces that are closed (they have no boundary) and compact (they do not in any sense go off to infinity). Other surfaces include the two-holed torus, and more generally the n -holed torus (the surfaces of genus 2, 3, 4, ...). To create these in a similar way, we need *Fuchsian groups*.

It is natural to expect that we can get other surfaces by using polygons with more than four sides. It turns out that if you use a polygon with eight sides, for example a regular octagon, and glue sides 1 and 3 together, 2 and 4 together, 5 and 7 together, and 6 and 8 together, you get the two-holed torus. How can we use a group to achieve the same result, as we did with the torus? For that we need a way of fitting lots of copies of the octagon together so that they overlap only along edges. The problem is that one cannot tile the plane with octagons: the angles of an octagon are 135° , and that is far too big because we need eight octagons to fit together at each vertex.

The way forward here is to use HYPERBOLIC GEOMETRY [I.3 §6.6] instead of Euclidean geometry. But we can also work with our bare hands. Take the unit disk in the complex plane, $\mathbb{D} = \{z : |z| \leq 1\}$. Take the group of what are called *Möbius transformations*, which are maps of the form $z \mapsto (az + b)/(cz + d)$. It is a routine calculation to show that these maps send circles and straight lines to circles and straight lines (they mix the two types up, sometimes sending a circle to a

straight line and vice versa) and that they map angles to equal but opposite angles, just like the more familiar Euclidean reflections. If we now select just those Möbius transformations that map \mathbb{D} to itself, then we have a group that we shall call G . Indeed, we very nearly have a Fuchsian group.

We need to find a shape that will play the role that the square played in the Euclidean plane. Our group G has the property that it maps diameters of \mathbb{D} and arcs of circles perpendicular to the boundary of \mathbb{D} to diameters of \mathbb{D} and arcs of circles perpendicular to the boundary of \mathbb{D} , so we let these play the role of straight lines and use eight of them as the edges of a (non-Euclidean) octagon. We find that we can do this in many ways, so we pick one with the highest degree of symmetry to make things easy for ourselves. That is, we draw a “regular octagon” centered on the center of the disk \mathbb{D} . This still leaves us with some choice: the bigger the octagon, the smaller its angles. So we draw the octagon with angles of $\pi/4$, which allows eight of them to cluster at each vertex, and then we can fit them together as we want. If we identify points that lie in corresponding places in different copies of the polygon, then the resulting space is a RIEMANN SURFACE [III.81] of genus 2.

A Fuchsian group is a subgroup of the group G (of Möbius transformations that map \mathbb{D} to itself) that moves some polygon around “en bloc” and thereby tiles the disk. Just as with the torus, we have a notion of equivalent points (ones that are in the corresponding place in different tiles) and when we identify equivalent points we get the space that we would also have obtained by identifying the edges of the polygon in pairs, which is the space we wanted.

All this can be described in the language of hyperbolic geometry. The *disk model* is defined by means of a RIEMANNIAN METRIC [I.3 §6.10] on \mathbb{D} , the differential of which is given by

$$ds = \frac{|dz|}{\sqrt{1 - |z|^2}}.$$

The elements of G move figures around in \mathbb{D} in a way that preserves hyperbolic distances. It follows that the geometry on the surface that we obtain by identifying points in the manner just described is *locally hyperbolic*, just as that of the torus was locally Euclidean.

It turns out that if we carry out the above construction starting with a regular $4n$ -sided figure (with $n > 2$), then we obtain a Riemann surface of genus n . But mathematicians can do much more. If you go back to the

plane and start not with a square but with a rectangle, or still more generally a parallelogram, it is reasonably easy to see that the same construction can be carried out. Indeed, if you just watch the original construction from an appropriate angle, instead of from vertically above the plane, then the square will turn into any parallelogram you choose (possibly enlarged or contracted). When you use a parallelogram, you again obtain a torus, but it differs from the original one in the same way that the square and the parallelogram differ: angles are distorted. It is a not entirely trivial exercise to show that the only angle-preserving maps from one parallelogram to another are similarities (uniform scaling by the same amount in two, and therefore all, directions). So the resulting tori have a different sense of what angles are: that is, they have different *conformal structures*.

The same happens in the hyperbolic disk. If one picks a $4n$ -sided polygon (its sides are parts of geodesics) whose edges come in pairs of equal length, and one finds a group that moves this polygon around en bloc and matches the edges exactly, then a Riemann surface is once again obtained, but if the polygons are not conformally equivalent, then neither are the corresponding surfaces; they have the same genus, n , but different conformal structures. We can even go further and allow some of the vertices of the polygon to lie on the boundary of the disk, in which case the corresponding sides of the polygon are infinitely long with respect to the hyperbolic metric. The space we then construct is a “punctured” Riemann surface, and again mathematicians can vary its conformal structure.

The fundamental importance of Fuchsian groups derives from the *uniformization theorem*, which says that all but the simplest Riemann surfaces arise from some Fuchsian group in the fashion described above. This includes every Riemann surface of genus greater than 1, and those of genus 1 with at least one puncture, with any possible conformal structure.

The name Fuchsian group was given to these groups by POINCARÉ [VI.61] in 1881, who discovered them in the course of work on the hypergeometric equation and related differential equations, which had been inspired by the work of the German mathematician Lazarus Fuchs. KLEIN [VI.57] protested to him that a better procedure might have been to name them after Schwarz, and Poincaré was willing to agree once he read the relevant paper by Schwarz, but by then Fuchs had given his approval to the name. When Klein protested too much (in Poincaré’s view), Poincaré publicly gave the name

PUP: I can confirm that this sentence is indeed accurate - in hyperbolic geometry!

Kleinian groups to the analogous class of groups that arise in the study of conformal transformations of the three-dimensional unit ball. The names have stuck ever since, but the study of Kleinian groups is much more difficult and neither Poincaré nor Klein could do much with the concept. However, the idea that every Riemann surface might arise from either the sphere, the Euclidean plane, or the hyperbolic plane was something they both came to conjecture. Rigorous proofs of this statement, the uniformization theorem, were to be given only in 1907, by Poincaré and Koebe independently.

The formal definition of a Fuchsian group is as follows. A subgroup H of the group of all Möbius transformations is said to act *discontinuously* if, for every compact set K in the disk \mathbb{D} the sets $h(K)$ and K are disjoint except for finitely many $h \in H$. A *Fuchsian group* is a subgroup H of the group of all Möbius transformations that acts discontinuously on the disk \mathbb{D} .

III.29 Function Spaces

Terence Tao

1 What Is a Function Space?

When one works with real or complex numbers, there is a natural notion of the *magnitude* of a number x , namely its modulus $|x|$. One can also use this notion of magnitude to define a distance $|x - y|$ between two numbers x and y and thereby say in a quantitative way which pairs of numbers are close and which ones are far apart.

The situation becomes more complicated, however, when one deals with objects with more degrees of freedom. Consider for instance the problem of determining the “magnitude” of a three-dimensional rectangular box. There are several candidates for such a magnitude: length, width, height, volume, surface area, diameter (the length of a long diagonal), eccentricity, and so forth. Unfortunately, these magnitudes do not give equivalent comparisons: for example, box A may be longer and have a greater volume than box B, but box B may be wider and have a greater surface area. Because of this, one abandons the idea that there should be only one notion of “magnitude” for boxes, and instead accepts that there is a multiplicity of such notions and that they can all be useful: for some applications one may wish to distinguish the large-volume boxes from the small-volume boxes, while in others one may wish to distinguish the eccentric boxes from the round boxes. Of course, there are several relationships

between the different notions of magnitude (e.g., the ISOPERIMETRIC INEQUALITY [IV.24] allows one to place an upper limit on the possible volume if one knows the surface area), so the situation is not as disorganized as it may at first appear.

Now let us turn to functions with a fixed domain and range. (A good case to have in mind is functions $f : [-1, 1] \rightarrow \mathbb{R}$ from the interval $[-1, 1]$ to the real line \mathbb{R} .) These objects have infinitely many degrees of freedom, so it should not be surprising that there are now infinitely many distinct notions of “magnitude,” which all provide different answers to the question “how large is a given function f ?” (or to the closely related question “how close together are two functions f and g ?”). In some cases, certain functions may have infinite magnitude by one measure and finite magnitude by another (similarly, a pair of functions may be very close by one measure and very far apart by another). Again, this situation may seem chaotic, but it simply reflects the fact that functions have many distinct characteristics—some are tall, some are broad, some are smooth, some are oscillatory, and so forth—and that, depending on the application at hand, one may need to give more weight to one of these characteristics than to others. In analysis, these characteristics are embodied in a variety of standard *function spaces* and their associated *norms*, which are available to describe functions both qualitatively and quantitatively.

Formally, a function space is a NORMED SPACE [III.64] X , the elements of which are functions (with some fixed domain and range). A majority (but certainly not all) of the standard function spaces considered in analysis are not just normed spaces but also BANACH SPACES [III.64]. The norm $\|f\|_X$ of a function f in X is the function space’s way of measuring how large f is. It is common, though not universal, for the norm to be defined by a simple formula and for the space X to consist precisely of those functions f for which the resulting definition $\|f\|_X$ makes sense and is finite. Thus, the mere fact that a function f belongs to a function space X can already convey some qualitative information about that function. For example, it may imply some regularity,¹ decay, boundedness, or integrability on the function f . The actual value of the norm $\|f\|_X$ makes this information quantitative. It may tell us *how* regular f is, *how much* decay it has, *by which constant* it is bounded, or *how large* its integral is.

1. The more smoothly a function varies, the more “regular” it is considered to be.

T&T note: check that we don’t have a run of five end-of-line hyphens here before CRC.

2 Examples of Function Spaces

We now present a sample of commonly used function spaces. For simplicity we shall consider only spaces of functions from $[-1, 1]$ to \mathbb{R} .

2.1 $C^0[-1, 1]$

This is the space of all CONTINUOUS FUNCTIONS [I.3 §5.2] from $[-1, 1]$ to \mathbb{R} , and is sometimes denoted $C[-1, 1]$. Continuous functions are regular enough to allow one to avoid many of the technical subtleties associated with very rough functions. Continuous functions on a COMPACT [III.9] interval such as $[-1, 1]$ are bounded, so the most natural norm to place on this space is the *supremum norm*, denoted $\|f\|_\infty$, which is the largest possible value of $|f(x)|$. (Formally, it is defined to be $\sup\{|f(x)| : x \in [-1, 1]\}$, but for continuous functions on $[-1, 1]$ the two definitions are equivalent.)

The supremum norm is the norm associated with uniform convergence: a sequence f_1, f_2, \dots converges uniformly to f if and only if $\|f_n - f\|_\infty$ tends to 0 as n tends to ∞ . The space $C^0[-1, 1]$ has the useful property that one can multiply functions together as well as adding them. This makes it a basic example of a *Banach algebra*.

2.2 $C^1[-1, 1]$

This is a space that has a more restricted membership than $C^0[-1, 1]$: not only must a function f in $C^1[-1, 1]$ be continuous but it must also have a derivative that is continuous. The supremum norm here is no longer a natural one, because a sequence of continuously differentiable functions can converge in this norm to a nondifferentiable function. Instead, the right norm here is the C^1 -norm $\|f\|_{C^1[-1, 1]}$, which is defined to be $\|f\|_\infty + \|f'\|_\infty$.

Notice that the C^1 -norm measures both the size of a function *and* the size of its derivative. (Merely controlling the latter would be unsatisfactory, since it would give constant functions a norm of zero.) Thus it is a norm that forces a greater degree of regularity than the supremum norm. One can similarly define the space $C^2[-1, 1]$ of twice continuously differentiable functions, and so forth, all the way up to the space $C^\infty[-1, 1]$ of infinitely differentiable functions. (There are also “fractional” versions of these spaces, such as $C^{0,\alpha}[-1, 1]$, the space of α -Hölder continuous functions. We will not discuss these variants here.)

2.3 The Lebesgue Spaces $L^p[-1, 1]$

The supremum norm $\|f\|_\infty$ mentioned earlier gives simultaneous control on the size of $|f(x)|$ for all $x \in [-1, 1]$. However, this means that if there is a tiny set of x for which $|f(x)|$ is very large, then $\|f\|_\infty$ is very large, even if a typical value of $|f(x)|$ is much smaller. It is sometimes more advantageous to work with norms that are less influenced by the values of a function on small sets. The L^p -norm of a function f is

$$\|f\|_p = \left(\int_{-1}^1 |f(x)|^p dx \right)^{1/p}.$$

This is defined for $1 \leq p < \infty$ and for any measurable f . The function space $L^p[-1, 1]$ is the class of measurable functions for which the above norm is finite. The norm $\|f\|_\infty$ of a measurable function f is its *essential supremum*: roughly speaking this means the largest value of $|f(x)|$ if you ignore sets of measure zero. It turns out to be the limit of the norms $\|f\|_p$ as p tends to infinity. The space $L^\infty[-1, 1]$ consists of those measurable functions f for which $\|f\|_\infty$ is finite. While the L^∞ norm is concerned solely with the “height” of a function, the L^p norms are instead concerned with a combination of the “height” and “width” of a function.

Particularly important among these norms is the L^2 -norm, since $L^2[-1, 1]$ is a HILBERT SPACE [III.37]. This space is exceptionally rich in symmetries: there is a wide variety of *unitary transformations*, that is, invertible linear maps T defined on $L^2[-1, 1]$ such that $\|Tf\|_2 = \|f\|_2$ for every function $f \in L^2[-1, 1]$.

2.4 The Sobolev Spaces $W^{k,p}[-1, 1]$

The Lebesgue norms control, to some extent, the height and width of a function, but say nothing about regularity; there is no reason why a function in L^p should be differentiable or even continuous. To incorporate such information one often turns to the *Sobolev norms* $\|f\|_{W^{k,p}[-1, 1]}$, defined for $1 \leq p \leq \infty$ and $k \geq 0$ by

$$\|f\|_{W^{k,p}[-1, 1]} = \sum_{j=0}^k \left\| \frac{d^j f}{dx^j} \right\|_p.$$

The *Sobolev space* $W^{k,p}[-1, 1]$ is the space of functions for which this norm is finite. Thus, a function lies in $W^{k,p}[-1, 1]$ if it and its first k derivatives all belong to $L^p[-1, 1]$. There is one subtlety: we do not require f to be k times differentiable in the usual sense, but in the weaker sense of DISTRIBUTIONS [III.18]. For instance, the function $f(x) = |x|$ is not differentiable at zero, but it does have a natural weak derivative: the function $f'(x)$ which is -1 when $x < 0$ and $+1$ when $x > 0$.

This function lies in $L^\infty[-1, 1]$ (since the set $\{0\}$ has measure zero, we do not need to specify $f'(0)$), and therefore f lies in $W^{1,\infty}[-1, 1]$ (which turns out to be the space of *Lipschitz-continuous* functions). We need to consider these generalized differentiable functions because without them the space $W^{k,p}[-1, 1]$ would not be complete.

Sobolev norms are particularly natural and useful in the analytical study of partial differential equations and mathematical physics. For instance, the $W^{1,2}$ norm can be interpreted as (the square root of) an “energy” associated with a function.

3 Properties of Function Spaces

There are many ways in which knowledge of the structure of function spaces can assist in the study of functions. For instance, if one has a good basis for the function space, so that every function in the space is a (possibly infinite) linear combination of basis elements, and one has some quantitative estimates on how this linear combination converges to the original function, then this allows one to represent that function efficiently in terms of a number of coefficients, and also allows one to approximate that function by smoother functions. For instance, one basic result about $L^2[-1, 1]$ is the *Plancherel theorem*, which asserts, among other things, that there are numbers $(a_n)_{n=-\infty}^\infty$ such that

$$\left\| f - \sum_{n=-N}^N a_n e^{\pi i n x} \right\|_2 \rightarrow 0 \quad \text{as } N \rightarrow \infty.$$

This shows that any function in $L^2[-1, 1]$ can be approximated to any desired accuracy in L^2 by a *trigonometric polynomial*: that is, an expression of the form $\sum_{n=-N}^N a_n e^{\pi i n x}$. The number a_n is the n th *Fourier coefficient* $\hat{f}(n)$ of f . It is given by the formula

$$\hat{f}(n) = \frac{1}{2} \int_{-1}^1 f(x) e^{-\pi i n x} dx.$$

One can regard this result as saying that the functions $e^{\pi i n x}$ form a very good basis for $L^2[-1, 1]$. (They are in fact an *orthonormal basis*: they have norm 1 and the inner product of two different ones is always zero.)

Another very basic fact about function spaces is that certain function spaces embed into others, so that a function from one space automatically also belongs to other spaces. Furthermore, there is often some inequality that gives an upper bound for one norm in terms of another. For instance, a function in a high-regularity space such as $C^1[-1, 1]$ automatically belongs to a low-regularity space such as $C^0[-1, 1]$,

and a function in a high-integrability space such as $L^\infty[-1, 1]$ automatically belongs to a low-integrability space such as $L^1[-1, 1]$. (This statement is no longer true if one replaces the interval $[-1, 1]$ by a set of infinite measure, such as the real line \mathbb{R} .) These inclusions cannot be reversed; however, one does have the *Sobolev embedding theorem*, which allows one to “trade” regularity for integrability. This result tells us that spaces with lots of regularity but low integrability can be embedded into spaces with low regularity but high integrability. A sample estimate of this type is

$$\|f\|_\infty \leq \|f\|_{W^{1,1}[-1,1]},$$

which tells us that if the integrals of $|f(x)|$ and $|f'(x)|$ are both finite, then f must be bounded (which is a far stronger integrability condition than the finiteness of $\|f\|_1$).

Another very useful concept is that of *DUALITY* [III.19]. Given a function space X , one can define the dual space X^* , which is formally defined as the class of all *continuous linear functionals* on X , or more precisely all maps $\omega : X \rightarrow \mathbb{R}$ (or $\omega : X \rightarrow \mathbb{C}$, if the function space is complex valued) that are linear and continuous with respect to the norm of X . For example, it turns out that every linear functional ω on the space $L^p[-1, 1]$ is of the form

$$\omega(f) = \int_{-1}^1 f(x) g(x) dx$$

for some function g in $L^q[-1, 1]$, where q is the *dual* or *conjugate exponent* of p , defined by the equation $1/p + 1/q = 1$.

One can sometimes analyze functions in a function space by looking instead at how the linear functionals in the dual space act on those functions. Similarly, one can often analyze a continuous linear operator $T : X \rightarrow Y$ from one function space to another by first considering the *adjoint operator* $T^* : Y^* \rightarrow X^*$, defined for all linear functionals $\omega : Y \rightarrow \mathbb{R}$ by letting $T^*\omega$ be the functional on X defined by the formula $T^*\omega(x) = \omega(Tx)$.

We mention one more important fact about function spaces, which is that certain function spaces X “interpolate” between two other function spaces X_0 and X_1 . For example, there is a natural sense in which the spaces $L^p[-1, 1]$ with $1 < p < \infty$ “lie between” the spaces $L^1[-1, 1]$ and $L^\infty[-1, 1]$. The precise definition of interpolation is too technical for this article, but its usefulness lies in the fact that the “extreme” spaces X_0 and X_1 are often easier to deal with than the “intermediate” spaces X . For this reason, it is sometimes

possible to prove difficult results about X by proving much easier results about X_0 and X_1 and “interpolating” between them. For instance, it can be used to give a short proof of *Young’s inequality*, which is the following statement. Let $1 \leq p, q, r \leq \infty$ satisfy the equation $1/p + 1/q = 1/r + 1$, let f and g belong to $L^p(\mathbb{R})$ and $L^q(\mathbb{R})$, respectively, and let $f * g$ be the *convolution* of f and g : that is, $f * g(x) = \int_{-\infty}^{\infty} f(y)g(x - y) dy$. Then

$$\begin{aligned} & \left(\int_{-\infty}^{\infty} |f * g(x)|^r dx \right)^{1/r} \\ & \leq \left(\int_{-\infty}^{\infty} |f(x)|^p dx \right)^{1/p} \left(\int_{-\infty}^{\infty} |g(x)|^q dx \right)^{1/q}. \end{aligned}$$

Interpolation is useful here because the inequality is easy to prove in the extreme cases when $p = 1$, when $q = 1$, or when $r = \infty$. It is much harder to prove this result without the help of interpolation theory.

III.30 Galois Groups

Given a polynomial function f , the *splitting field* of f is defined to be the smallest FIELD [I.3 §2.2] that contains all rational numbers and all the roots of f . The *Galois group* of f is the group of all AUTOMORPHISMS [I.3 §4.1] of the splitting field. Each such automorphism permutes the roots of f , so the Galois group can be thought of as a subset of the group of all PERMUTATIONS [III.70] of these roots. The structure and properties of the Galois group are closely connected with the solubility of the polynomial: in particular, the Galois group can be used to show that not all polynomials are *solvable by radicals* (that is, solvable by means of a formula that involves the usual arithmetic operations together with the extraction of roots). This theorem, spectacular as it is, is by no means the only application of Galois groups: they play a central role in modern algebraic number theory.

For more details, see THE INSOLUBILITY OF THE QUINTIC [V.24] and ALGEBRAIC NUMBERS [IV.3 §20].

III.31 The Gamma Function

Ben Green

If n is a positive integer, then its *factorial*, written $n!$, is the number $1 \times 2 \times \cdots \times n$: that is, the product of all positive integers up to n . For example, the first eight factorials are 1, 2, 6, 24, 120, 720, 5040, and 40320. (The exclamation mark was introduced by Christian Kramp 200 years ago as a convenience to the printer: it is perhaps also intended to convey some alarm at

the rapidity with which $n!$ grows. An obsolete notation, which can still be found in some twentieth-century texts, is \underline{n} .) From this definition, it might appear to be impossible to make sense of the idea of the factorial of a number that is not a positive integer, but, as it turns out, it is not just possible to do so, but also extremely useful.

The *gamma function*, written Γ , is a function that agrees with the factorial function at positive integer values, but that makes sense for any real number, and even for any complex number. Actually, for various reasons it is natural to define Γ so that $\Gamma(n) = (n-1)!$ for $n = 2, 3, \dots$. Let us start by writing

$$\Gamma(s) = \int_0^{\infty} x^{s-1} e^{-x} dx, \quad (1)$$

without paying too much attention to whether the integral converges. If we integrate by parts, then we find that

$$\Gamma(s) = [-x^{s-1} e^{-x}]_0^{\infty} + \int_0^{\infty} (s-1)x^{s-2} e^{-x} dx. \quad (2)$$

As x tends to infinity, $x^{s-1} e^{-x}$ tends to zero, and if s is, for example, a real number greater than 1, then $x^{s-1} = 0$ when $x = 0$. Therefore, for such s , we can ignore the first term in the above expression. But the second one is simply the formula for $\Gamma(s-1)$, so we have shown that $\Gamma(s) = (s-1)\Gamma(s-1)$, which is just what we need if we want to think of $\Gamma(s)$ as something like $(s-1)!$.

It is not hard to show that the integral is in fact convergent whenever s is a *complex* number and $\operatorname{Re}(s)$ (the real part of s) is positive. Moreover, it defines a HOLOMORPHIC FUNCTION [I.3 §5.6] in that region. When the real part of s is negative, the integral does not converge at all, and so the formula (1) cannot be used to define the gamma function in its entirety. However, we can instead use the property $\Gamma(s) = (s-1)\Gamma(s-1)$ to *extend* the definition. For example, when $-1 < \operatorname{Re}(s) \leq 0$, we know that the definition does not work directly, but it does work for $s+1$, since $\operatorname{Re}(s+1) > 0$. We would like $\Gamma(s+1)$ to equal $s\Gamma(s)$, so it makes sense to define $\Gamma(s)$ to be $\Gamma(s+1)/s$. Once we have done this, we can turn our attention to values of s with $-2 < \operatorname{Re}(s) \leq -1$, and so on.

The reader may object that in defining $\Gamma(0)$ (for example), we have divided by zero. This is perfectly permissible, however, if all we require of Γ is that it should be MEROMORPHIC [V.34], because meromorphic functions are allowed to take the “value” ∞ . Indeed, it is not hard to see that Γ , as we have defined it, has simple poles at $0, -1, -2, \dots$.

There are in fact many functions that share the useful properties of Γ . (For instance, because $\cos(2\pi s) = \cos(2\pi(s+1))$ for any s , and $\cos(2\pi n) = 1$ for every integer n , the function $F(s) = \Gamma(s) \cos(2\pi s)$ also has the property $F(s) = (s-1)F(s-1)$ and $F(n) = (n-1)!$.) Nevertheless, for a variety of reasons, the function Γ , as we have defined it, is the most natural meromorphic extension of the factorial function. The most persuasive reason is the fact that it arises so often in natural contexts, but it is also, in a certain sense, the smoothest interpolation of the factorial function to all positive real values. In fact, if $f : (0, \infty) \rightarrow (0, \infty)$ is such that $f(x+1) = xf(x)$, $f(1) = 1$, and $\log f$ is convex, then $f = \Gamma$.

There are many interesting formulas involving Γ , such as $\Gamma(s)\Gamma(1-s) = \pi/\sin(\pi s)$. There is also the famous result $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, which is essentially equivalent to the fact that the area under the “normal distribution curve” $h(x) = (1/\sqrt{2\pi})e^{-x^2/2}$ is 1 (this can be seen by making the substitution $x = u^2/2$ in (1)). A very important result concerning Γ is the Weierstrass product expansion, which states that

$$\frac{1}{\Gamma(z)} = ze^{yz} \prod_{n=1}^{\infty} \left(1 + \frac{z}{n}\right) e^{-z/n}$$

for all complex z , where γ is Euler’s constant:

$$\gamma = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{2} + \cdots + \frac{1}{n} - \log n\right).$$

This formula makes it clear that Γ never vanishes, and that it has simple poles at 0 and the negative integers.

Why is the gamma function important? Reason enough is its frequent occurrence in many parts of mathematics, but one can attempt to explain why this should be so. One reason is that Γ , as defined in (1), is the *Mellin transform* of the unarguably natural function $f(x) = e^{-x}$. The Mellin transform is a type of FOURIER TRANSFORM [III.27], but it is defined for functions on the group (\mathbb{R}^+, \times) rather than $(\mathbb{R}, +)$ (which is the habitat of the most familiar type of Fourier transform). For this reason, Γ is often seen in number theory, particularly ANALYTIC NUMBER THEORY [IV.4], where multiplicatively defined functions are often studied by taking Fourier transforms.

One appearance of Γ in a number-theoretical context is in the functional equation for the RIEMANN ZETA FUNCTION [IV.4 §3], namely,

$$\Xi(s) = \Xi(1-s),$$

where

$$\Xi(s) = \Gamma(s/2)\pi^{-s/2}\zeta(s). \quad (3)$$

The ζ function has a well-known product representation

$$\zeta(s) = \prod_p (1 - p^{-s})^{-1},$$

where the product is over primes and the representation is valid for $\text{Re}(s) > 1$. The extra factor $\Gamma(s/2)\pi^{-s/2}$ in (3) may be regarded as coming from the “prime at infinity” (a term which may be rigorously defined).

Stirling’s formula is a very useful tool in dealing with the gamma function: it provides a rather accurate estimate for $\Gamma(z)$ in terms of simpler functions. A very rough (but often useful) approximation for $n!$ is $(n/e)^n$, which tells us that $\log(n!)$ is about $n(\log n - 1)$. Stirling’s formula is a sharper version of this crude estimate. Let $\delta > 0$ and suppose that z is a complex number that has modulus at least 1 and argument between $-\pi + \delta$ and $\pi - \delta$. (This second condition keeps z away from the negative real axis, where the poles are.) Then Stirling’s formula states that

$$\log \Gamma(z) = (z - \frac{1}{2}) \log z - z + \frac{1}{2} \log 2\pi + E,$$

where the error E is at most $C(\delta)/|z|$. Here, $C(\delta)$ stands for a certain positive real number that depends on δ . (The smaller you make δ , the larger you have to make $C(\delta)$.) Using this, one may confirm that Γ decays exponentially as $\text{Im } z \rightarrow \infty$ in any fixed vertical strip in the complex plane. In fact, if $\alpha < \sigma < \beta$, then

$$|\Gamma(\sigma + it)| \leq C(\alpha, \beta) |t|^{\beta-1} e^{-\pi|t|/2}$$

for all $|t| > 1$, uniformly in σ .

III.32 Generating Functions

Suppose that you have defined a combinatorial structure, and for each nonnegative integer n you wish to understand how many examples of this structure there are of size n . If a_n denotes this number, then the object that you are trying to analyze is the sequence $a_0, a_1, a_2, a_3, \dots$. If the structure is quite complicated, then this may be a very hard problem, but one can sometimes make it easier by considering a different object, the *generating function* of the sequence, which contains the same information.

To define this function, one simply regards the sequence a_n as the sequence of coefficients in a power series. That is, the generating function f of the sequence is given by the formula

$$f(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots$$

The reason this can be useful is that one can sometimes derive a succinct expression for f and analyze it

without reference to the individual numbers a_n . For example, one important generating function has the formula $f(x) = (1 - \sqrt{1 - 4x})/2x$. In such cases, one can deduce properties of the sequence a_0, a_1, a_2, \dots from properties of f , rather than the other way round.

For more on generating functions, see ENUMERATIVE AND ALGEBRAIC COMBINATORICS [IV.22] and TRANSFORMS [III.93].

III.33 Genus

The *genus* is a topological invariant of surfaces: that is, a quantity associated with a surface that does not change when the surface is continuously deformed. Roughly speaking, it corresponds to the number of holes of that surface, so a sphere has genus 0, a torus has genus 1, a pretzel shape (that is, the surface of a blown-up figure of eight) has genus 2, and so on. If one triangulates an orientable surface and counts the number of vertices, edges, and faces in the triangulation, denoting them V , E , and F , respectively, then the *Euler characteristic* is defined to be $V - E + F$. It can be shown that if g is the genus and χ is the Euler characteristic, then $\chi = 2 - 2g$. See [I.4 §2.2] for a fuller discussion.

A famous result of POINCARÉ [VI.61] states that for every nonnegative integer g there is precisely one orientable surface of genus g . (Moreover, genus can also be defined for nonorientable surfaces, where a similar result holds.) See DIFFERENTIAL TOPOLOGY [IV.9 §2.3] for more about this theorem.

One can associate an orientable surface, and therefore a genus, with a smooth algebraic curve. An ELLIPTIC CURVE [III.21] can be defined as a smooth curve of genus 1. See ALGEBRAIC GEOMETRY [IV.7 §10] for more details.

III.34 Graphs

A graph is one of the simplest of all mathematical structures: it consists of some elements called *vertices* (of which there are usually just finitely many), some pairs of which are deemed to be “joined” or “adjacent.” It is customary to represent the vertices by points in a plane and to join adjacent points by a line. The line is referred to as an *edge* (though how the line is drawn or visualized is irrelevant: all that is important is whether or not two points are joined).

For example, the rail network of a country can be represented by a graph: we can use vertices to represent

the stations, and we can join two vertices if they represent consecutive stations along some rail line. Another example Another example is provided by the Internet: the vertices are all the world’s computers, and two are adjacent if there is a direct link between them.

Many questions in graph theory take the form of asking what some structural property of graphs can tell you about its other properties. For example, suppose that we are trying to find a graph with n vertices that does not contain a triangle (defined to be a set of three vertices that are mutually joined). How many edges can the graph have? Clearly $\frac{1}{4}n^2$ is possible, at least if n is even, since one can then divide up the n vertices into two equal classes and join all vertices in one class to all vertices in the other. But can there be more edges than that?

Here is another example of a typical question about graphs. Let k be a positive integer. Must there exist an n such that every graph with n vertices always contains either k vertices that are all joined to each other or k vertices none of which are joined to each other? This question is quite easy for $k = 3$ (where $n = 6$ suffices), but already for $k = 4$ it is not obvious that such an n exists.

For more on these problems (the first is the founding problem of “extremal graph theory,” while the second is the founding problem of “Ramsey theory”) and on the study of graphs in general, see EXTREMAL AND PROBABILISTIC COMBINATORICS [IV.23].

III.35 Hamiltonians

Terence Tao

At first glance, the many theories and equations of modern physics exhibit a bewildering diversity: compare, for instance, classical mechanics with quantum mechanics, nonrelativistic physics with relativistic physics, or particle physics with statistical mechanics. However, there are strong unifying themes connecting all of these theories. One of these is the remarkable fact that in all of them the evolution of a physical system over time (as well as the steady states of that system) is largely controlled by a single object, the *Hamiltonian* of that system, which can often be interpreted as describing the total energy of any given state in that system. Roughly speaking, each physical phenomenon (e.g., electromagnetism, atomic bonding, particles in a potential well, etc.) may correspond to a single Hamiltonian H , while each type of mechanics (classical, quantum, statistical, etc.) corresponds to a different way

Query for PUP:
how would an
American say this?

of using that Hamiltonian to describe a physical system. For instance, in classical physics, the Hamiltonian is a function $(q, p) \mapsto H(q, p)$ of the positions q and momenta p of the system, which then evolve according to Hamilton's equations:

$$\frac{dq}{dt} = \frac{\partial H}{\partial p}, \quad \frac{dp}{dt} = -\frac{\partial H}{\partial q}.$$

In (nonrelativistic) quantum mechanics, the Hamiltonian H becomes a LINEAR OPERATOR [III.52] (which is often a formal combination of the position operators q and momenta operators p), and the wave function ψ of the system then evolves according to THE SCHRÖDINGER EQUATION [III.85]:

$$i\hbar \frac{d}{dt} \psi = H\psi.$$

In statistical mechanics, the Hamiltonian H is a function of the microscopic state (or *microstate*) of a system, and the probability that a system at a given temperature T will lie in a given microstate is proportional to $e^{-H/kT}$. And so on and so forth.

Many fields of mathematics are closely intertwined with their counterparts in physics, and so it is not surprising that the concept of a Hamiltonian also appears in pure mathematics. For instance, motivated by classical physics, Hamiltonians (as well as generalizations of Hamiltonians, such as *moment maps*) play a major role in dynamical systems, differential equations, Lie group theory, and symplectic geometry. Motivated by quantum mechanics, Hamiltonians (as well as generalizations, such as *observables* or *pseudo-differential operators*) are similarly prominent in operator algebras, spectral theory, representation theory, differential equations, and microlocal analysis.

Because of their presence in so many areas of physics and mathematics, Hamiltonians are useful for building bridges between seemingly unrelated fields: for instance, between classical mechanics and quantum mechanics, or between symplectic mechanics and operator algebras. The properties of a given Hamiltonian often reveal much about the physical or mathematical objects associated with that Hamiltonian. For example, the symmetries of a Hamiltonian often induce corresponding symmetries in objects described using that Hamiltonian. While not every interesting feature of a mathematical or physical object can be read off directly from its Hamiltonian, this concept is still fundamental to understanding the properties and behavior of such objects.

See also VERTEX OPERATOR ALGEBRAS [IV.13 §2.1], MIRROR SYMMETRY [IV.14 §§2.1.3, 2.2.1], and SYMPLECTIC MANIFOLDS [III.90 §2.1].

III.36 The Heat Equation

Igor Rodnianski

The heat equation was first proposed by FOURIER [VI.25] as a mathematical description of the transfer of heat in solid bodies. Its influence has subsequently been felt in many corners of mathematics: it provides explanations for such disparate phenomena as the formation of ice (the *Stefan problem*), the theory of incompressible viscous fluids (the NAVIER-STOKES EQUATION [III.23]), geometric flows (e.g., curve shortening, and the harmonic-map heat flow problem), BROWNIAN MOTION [IV.25], liquid filtration in porous media (the *Hele-Shaw problem*), index theorems (e.g., the *Gauss-Bonnet-Chern formula*), the price of stock options (the BLACK-SCHOLES FORMULA [VII.9 §2]), and the topology of three-dimensional manifolds (THE POINCARÉ CONJECTURE [V.28]). But the bright future of the heat equation could have been predicted at its birth: after all, another small event that accompanied it was the creation of FOURIER ANALYSIS [III.27].

The propagation of heat is based on a simple continuity principle. The change in the quantity of heat u in a small volume ΔV over a small interval of time Δt is approximately

$$CD \frac{\partial u}{\partial t} \Delta t \Delta V,$$

where C is the heat capacity of the substance and D is its density; but it is also given by the amount of heat entering and exiting through ΔV , which is approximately

$$K \Delta t \int_{\partial \Delta V} \frac{\partial u}{\partial \mathbf{n}},$$

where K is the heat conductivity constant and \mathbf{n} is the unit normal to the boundary of ΔV .

Thus, setting the values of all physical constants to 1, dividing through by Δt and ΔV , and letting them tend to zero, we find that the evolution of the amount of heat (that is, the temperature) in a three-dimensional solid Ω is governed by the following classical heat equation, where $u(t, \mathbf{x})$ is the temperature at time t at the point $\mathbf{x} = (x, y, z)$:

$$\frac{\partial}{\partial t} u(t, \mathbf{x}) - \Delta u(t, \mathbf{x}) = 0. \quad (1)$$

Here

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

is the three-dimensional Laplacian; Δu is the limit as the diameter of ΔV tends to zero of the quantity

$$\frac{1}{\Delta V} \int_{\Delta V} \frac{\partial u}{\partial \mathbf{n}}.$$

To determine $u(t, \mathbf{x})$, equation (1) needs to be complemented by the *initial distribution* $u_0(\mathbf{x}) = u(0, \mathbf{x})$ and *boundary conditions* on the solid interface $\partial\Omega$. For example, for a solid unit cube C with surface maintained at zero temperature, the heat equation is considered as a problem with Dirichlet boundary conditions and, as was proposed by Fourier, $u(t, \mathbf{x})$ can be found by the method of separation of variables by expanding $u_0(\mathbf{x})$ into its Fourier series

$$u_0(x, y, z) = \sum_{k,m,l=0}^{\infty} C_{kml} \sin(\pi k x) \times \sin(\pi m y) \sin(\pi l z),$$

which leads to the solution

$$u(t, x, y, z) = \sum_{k,m,l=0}^{\infty} e^{-\pi^2(k^2+m^2+l^2)t} C_{kml} \sin(\pi k x) \times \sin(\pi m y) \sin(\pi l z).$$

This simple example already illuminates a fundamental property of the heat equation: the tendency of its solutions to converge to an equilibrium state. In this case it reflects a physically intuitive fact that the temperature $u(t, \mathbf{x})$ converges to the constant distribution

$$u^*(\mathbf{x}) = C_{000}.$$

Propagation of heat in an insulated body corresponds to the choice of the *Neumann* boundary conditions, in which the normal derivative of u (normal, that is, to the boundary $\partial\Omega$) is set to vanish. Its solutions can be constructed in a similar fashion.

The reason that Fourier analysis is intimately connected with the heat equation is that the trigonometric functions are EIGENFUNCTIONS [I.3 §4.3] of the Laplacian. A variety of more general heat equations can be obtained if one replaces the Laplacian by a more general linear, SELF-ADJOINT [III.52 §3.2], nonnegative HAMILTONIAN [III.35] H with a discrete set of eigenvalues λ_n and corresponding eigenfunctions ψ_n . That is, one considers the heat flow

$$\frac{\partial}{\partial t} u + H u = 0.$$

The solution $u(t)$ is given by the formula $u(t) = e^{-tH} u_0$, where e^{-tH} is the *heat semigroup* generated by H , which also takes the more explicit form

$$u(t, \mathbf{x}) = \sum_{n=0}^{\infty} e^{-\lambda_n t} C_n \psi_n(\mathbf{x}).$$

Here the coefficients C_n are the Fourier coefficients of u_0 relative to H : that is, they are the coefficients that arise when we write u_0 as a sum $\sum_{n=0}^{\infty} C_n \psi_n$. (The existence of such a decomposition follows from the SPECTRAL THEOREM [III.52 §3.4] for self-adjoint operators. In a similar way, heat flows can also be generated by self-adjoint operators with a continuous spectrum.) In particular, the asymptotic behavior of $u(t, \mathbf{x})$ as $t \rightarrow +\infty$ is completely determined by the spectrum of H .

Although explicit, representations like this do not provide very good quantitative descriptions of the behavior of the heat equation. To obtain such descriptions one has to abandon the idea of constructing solutions explicitly and look instead for principles and methods that apply to general classes of solutions while also being sufficiently robust to be useful in the analysis of more complicated heat equations.

The first methods of this type are called *energy identities*. To derive an energy identity, one multiplies the heat equation by a certain quantity, which may depend on the given solution, and integrates by parts. The simplest two identities of this type are the *conservation of total heat* of an insulated body,

$$\frac{d}{dt} \int_{\Omega} u(t, \mathbf{x}) d\mathbf{x} = 0,$$

and the energy identity,

$$\begin{aligned} \int_{\Omega} u^2(t, \mathbf{x}) d\mathbf{x} + 2 \int_0^t \int_{\Omega} |\nabla u(s, \mathbf{x})|^2 d\mathbf{x} ds \\ = \int_{\Omega} u^2(0, \mathbf{x}) d\mathbf{x}. \end{aligned}$$

The second identity already captures a fundamental smoothing property of the heat equation: since all three integrands are nonnegative and the first and third integrals are finite, the average of the mean-square gradient of u is finite, even if the initial mean-square gradient is infinite, and it even decreases to zero with t . In fact, away from the boundary of Ω an arbitrary amount of smoothing takes place, and not just on average but at every time $t > 0$.

The second fundamental principle of the heat equation is the *global maximum principle*

$$\begin{aligned} \max_{\mathbf{x} \in \Omega, 0 \leq t \leq T} u(t, \mathbf{x}) \\ \leq \max \left(u(0, \mathbf{x}), \max_{\mathbf{x} \in \partial\Omega, 0 \leq t \leq T} u(t, \mathbf{x}) \right), \end{aligned}$$

which tells us the familiar fact that the hottest spot in the body, over all time, is either on its boundary or in the initial distribution.

T&T note:
cross-references
for the spectral
theorem all need
to be checked at
the end of the
process as there
are a number of
places that they
could all point.

Finally, the diffusive properties of the heat equation in \mathbb{R}^n are captured by the *Harnack inequality* for nonnegative solutions u . It tells us that

$$\frac{u(t_2, \mathbf{x}_2)}{u(t_1, \mathbf{x}_1)} \geq \left(\frac{t_1}{t_2}\right)^{n/2} e^{-|\mathbf{x}_2 - \mathbf{x}_1|^2 / 4(t_2 - t_1)}$$

when $t_2 > t_1$. This tells us that if the temperature at \mathbf{x}_1 at time t_1 takes a certain value, then the temperature at \mathbf{x}_2 at time t_2 cannot be too much smaller.

This form of the Harnack inequality features a very important object in the study of the heat equation, called the *heat kernel*:

$$p(t, \mathbf{x}, \mathbf{y}) = \frac{1}{(4\pi t)^{n/2}} e^{-|\mathbf{x} - \mathbf{y}|^2 / 4t}.$$

One of its many uses is that it allows one to construct solutions of the heat equation in the whole of space (that is, in \mathbb{R}^n) from initial data u_0 , by the formula

$$u(t, \mathbf{x}) = \int_{\mathbb{R}^n} p(t, \mathbf{x}, \mathbf{y}) u_0(\mathbf{y}) d\mathbf{y}.$$

It also shows that after a time t initial point disturbances become distributed in a ball of radius \sqrt{t} around the point of the original disturbance. This sort of relation between spatial scales and timescales is the characteristic *parabolic scaling* of the heat equation.

As was shown by Einstein, the heat equation is intimately connected with the diffusion process of Brownian motion. In fact, the mathematical description of Brownian motion is in terms of a random process B_t with transitional probability densities given by the heat kernel $p(t, \mathbf{x}, \mathbf{y})$. For the n -dimensional Brownian motion $B_t^{\mathbf{x}}$ starting at \mathbf{x} , the function

$$u(t, \mathbf{x}) = \mathbb{E}[u_0(\sqrt{2}B_t^{\mathbf{x}})]$$

computed with the help of expectation value \mathbb{E} is precisely the solution of the heat equation in \mathbb{R}^n with initial data $u_0(\mathbf{x})$. This connection is the start of a mutually beneficial relationship between the theory of the heat equation and probability. Among the most profitable applications of this relationship is the *Feynman-Kac formula*

$$u(t, \mathbf{x}) = \mathbb{E} \left[\exp \left(- \int_0^t V(\sqrt{2}B_s^{\mathbf{x}}) ds \right) u_0(\sqrt{2}B_t^{\mathbf{x}}) \right],$$

which connects Brownian motion with solutions of the heat equation

$$\frac{\partial}{\partial t} u(t, \mathbf{x}) - \Delta u(t, \mathbf{x}) + V(\mathbf{x})u(t, \mathbf{x}) = 0$$

with initial data $u_0(\mathbf{x})$.

The three fundamental principles of the heat equation described above are remarkably robust, in the sense that they, or weaker versions of them, hold even

for very general variants of the classical equation. For instance, they can be applied to the question of the continuity of solutions of the heat equation

$$\frac{\partial}{\partial t} u - \sum_{i,j=1}^n \frac{\partial}{\partial x_i} \left(a_{ij}(\mathbf{x}) \frac{\partial}{\partial x_j} u \right) = 0,$$

where all that is assumed of the coefficients a_{ij} is that they are bounded and that they satisfy the *ellipticity condition* $\lambda|\xi|^2 \leq a_{ij}\xi^i\xi^j \leq \Lambda|\xi|^2$. One can even look at the equations in “nondivergence form”:

$$\frac{\partial}{\partial t} u - \sum_{i,j=1}^n a_{ij}(\mathbf{x}) \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_j} u = 0.$$

Here, the connection between the heat equation and the corresponding stochastic diffusion process turns out to be particularly helpful. This analysis has led to beautiful applications in the CALCULUS OF VARIATIONS [III.96] and in fully nonlinear problems.

The same principles also hold for the heat equations on RIEMANNIAN MANIFOLDS [I.3 §6.10]. The appropriate analogue of the Laplacian for a manifold M is the *Laplace-Beltrami operator* Δ_M , and the heat equation for M is

$$\frac{\partial}{\partial t} u - \Delta_M u = 0.$$

If the Riemannian metric is g , then in local coordinates Δ_M takes the form

$$\Delta_M = \frac{1}{\sqrt{\det g(\mathbf{x})}} \sum_{i,j=1}^n \frac{\partial}{\partial x_i} \left(g^{ij}(\mathbf{x}) \sqrt{\det g(\mathbf{x})} \frac{\partial}{\partial x_j} \right).$$

In this case, a version of the Harnack inequality holds for the heat equation on a manifold that has RICCI CURVATURE [III.80] bounded from below. Interest in the heat equations on manifolds is in part motivated by nonlinear geometric flows and attempts to understand their long-term behavior. One of the earliest geometric flows was the *harmonic map flow*

$$\frac{\partial}{\partial t} \Phi - \Delta_M^N \Phi = 0,$$

which describes a deformation of the map $\Phi(t, \cdot)$ between two compact Riemannian manifolds M and N . The operator Δ_M^N is a nonlinear Laplacian that is constructed by projecting Δ_M onto the tangent space of N . This is a *gradient flow* associated with the energy

$$E[U] = \frac{1}{2} \int_M |dU|_N^2;$$

it measures the stretching of the map U between M and N . Under the assumption that the *sectional curvature* of N is nonpositive, it can be shown that the harmonic map heat flow is regular and converges, as $t \rightarrow +\infty$, to

a harmonic map between M and N , which is a critical point of the energy functional $E[U]$. This heat equation is used to establish the existence of harmonic maps and to construct a continuous deformation of a given map $\Phi(0, \cdot)$ to a harmonic map $\Phi(+\infty, \cdot)$. The curvature assumption on the target manifold N is responsible for the crucial *monotonicity* properties of the harmonic map heat flow, which come to light through the use of the energy estimates.

An even more spectacular application of a deformation principle of this kind appears in the three-dimensional RICCI FLOW [III.80]

$$\frac{\partial}{\partial t} g_{ij} = -2\text{Ric}_{ij}(g),$$

which is a *quasilinear* heat evolution of a family of metrics $g_{ij}(t)$ on a given manifold M . In this case the flow is not necessarily regular; nonetheless, it can be extended as a flow with “surgeries” in such a way that the structure of the surgeries and the long-term behavior of the flow can be precisely analyzed. This analysis shows in particular that any three-dimensional simply connected manifold is diffeomorphic to a three-dimensional sphere, which gives the proof of the Poincaré conjecture.

The long-term behavior of the heat equation is also important in the analysis of *reaction-diffusion systems* and associated biological phenomena. This was suggested already in the work of TURING [VI.94] in his attempt to understand *morphogenesis* (the formation of inhomogeneous patterns such as animal-coat patterns from a nearly homogeneous initial state) by means of exponential instabilities in the reaction-diffusion equations

$$\frac{\partial}{\partial t} u = \mu \Delta u + f(u, v), \quad \frac{\partial}{\partial t} v = \nu \Delta v + g(u, v).$$

These examples emphasize the long-term behavior of the heat equation, and in particular the tendency of its solutions to converge to an equilibrium, or alternatively to develop exponential instabilities. However, it turns out that the short-term behavior of the heat equation on a manifold M is of the utmost importance in connection with the geometry and topology of M . This connection is twofold: first, one seeks to establish a relationship between the spectrum of Δ_M and the geometry of M ; second, one can use an analysis of the short-term behavior to prove *index theorems*. The former aspect, in the context of planar domains, is captured by Marc Kac’s well-known question, “Can one hear the shape of

a drum?” For manifolds it begins with the *Weyl formula*

$$\sum_{i=0}^{\infty} e^{-t\lambda_i} = \frac{1}{(4\pi t)^{n/2}} (\text{Vol}(M) + O(t))$$

as t tends to 0. The left-hand side of the identity is the trace of the heat kernel of Δ_M . That is,

$$\sum_{i=0}^{\infty} e^{-t\lambda_i} = \text{tr } e^{-t\Delta_M} = \int_M p(t, x, x) \, dx,$$

PUP: I can confirm that the repetition of x in ‘ (t, x, x) ’ is OK.

where $p(t, x, y)$ is such that any solution of the heat equation $\partial u / \partial t - \Delta_M u = 0$ with $u(0, x) = u_0(x)$ is given by the expression

$$u(t, x) = \int_M p(t, x, y) u_0(y) \, dy.$$

The right-hand side of the Weyl identity reflects the short-term asymptotics of the heat kernel $p(t, x, y)$.

The heat-flow approach to the proof of the index theorems can be viewed as a refinement of both sides of the Weyl identity. The trace on the left-hand side is replaced by a more complicated “super-trace,” while the right-hand side involves full asymptotics of the heat kernel, which requires one to understand subtle cancellations. The simplest example of this kind is the *Gauss-Bonnet formula*

$$\chi(M) = 2\pi \int_M R,$$

which connects the Euler characteristic of a two-dimensional manifold M and the integral of its scalar curvature. The Euler characteristic $\chi(M)$ arises from a linear combination of traces of the heat flows associated with the *Hodge Laplacian* $(d + d^*)^2$ restricted to the space of exterior differential 0-forms, 1-forms, and 2-forms. A proof of a general ATIYAH-SINGER INDEX THEOREM [V.2] involves heat flows associated with an operator given by the square of a *Dirac operator*.

III.37 Hilbert Spaces

The theory of VECTOR SPACES [I.3 §2.3] and LINEAR MAPS [I.3 §4.2] underpins a large part of mathematics. However, angles cannot be defined using vector space concepts alone, since linear maps do not in general preserve angles. An *inner product space* can be thought of as a vector space with just enough extra structure for the notion of angle to make sense.

The simplest example of an inner product on a vector space is the standard scalar product defined on \mathbb{R}^n , the space of all real sequences of length n , as follows. If $v = (v_1, \dots, v_n)$ and $w = (w_1, \dots, w_n)$ are two such sequences, then their scalar product, denoted $\langle v, w \rangle$,

is the sum $v_1 w_1 + v_2 w_2 + \cdots + v_n w_n$. (For example, the scalar product of $(3, 2, -1)$ and $(1, 4, 4)$ is $3 \times 1 + 2 \times 4 + (-1) \times 4 = 7$.)

Among the properties that the scalar product has are the following two.

- (i) It is linear in each variable separately. That is, $\langle \lambda u + \mu v, w \rangle = \lambda \langle u, w \rangle + \mu \langle v, w \rangle$ for any three vectors u, v , and w and any two scalars λ and μ , and similarly $\langle u, \lambda v + \mu w \rangle = \lambda \langle u, v \rangle + \mu \langle u, w \rangle$.
- (ii) The scalar product $\langle v, v \rangle$ of any vector v with itself is always a nonnegative real number, and is zero only if v is zero.

In a general vector space, any function $\langle v, w \rangle$ of pairs of vectors v and w that has these two properties is called an inner product, and a vector space with an inner product is called an inner product space. If the vector space has complex scalars, then instead of (i) one must use the following modification.

- (i') For any three vectors u, v , and w and any two scalars λ and μ , $\langle \lambda u + \mu v, w \rangle = \lambda \langle u, w \rangle + \mu \langle v, w \rangle$, and $\langle u, \lambda v + \mu w \rangle = \bar{\lambda} \langle u, v \rangle + \bar{\mu} \langle u, w \rangle$. That is, the inner product is *conjugate-linear* in the second variable.

The reason this has anything to do with angles is that in \mathbb{R}^2 and \mathbb{R}^3 the scalar product of two vectors v and w works out as the length of v times the length of w times the cosine of the angle between them. In particular, since a vector v makes an angle of zero with itself, $\langle v, v \rangle$ is the square of the length of v .

This gives us a natural way to *define* length and angle in an inner product space. The length, or *norm*, of a vector v , denoted $\|v\|$, is $\sqrt{\langle v, v \rangle}$. Given two vectors v and w , the angle between them is defined by the fact that it lies between 0 and π (or 180°) and its cosine is $\langle v, w \rangle / \|v\| \|w\|$. Once length has been defined, we can also talk about distance: the distance $d(v, w)$ between v and w is the length of their difference, or $\|v - w\|$. This definition of distance satisfies the axioms for a METRIC SPACE [III.58]. From the notion of angle, we can say what it is for v and w to be orthogonal to each other: this simply means that $\langle v, w \rangle = 0$.

The usefulness of inner product spaces goes far beyond their ability to represent the geometry of two- and three-dimensional space. Where they really come into their own is if they are infinite dimensional. Then it becomes convenient if they satisfy the additional property of *completeness*, which is briefly discussed at the

end of [III.64]. A complete inner product space is called a *Hilbert space*.

Two important examples of Hilbert spaces are the following.

- (i) ℓ_2 is the natural infinite-dimensional generalization of \mathbb{R}^n with the standard scalar product. It is the set of all infinite sequences (a_1, a_2, a_3, \dots) such that the infinite sum $|a_1|^2 + |a_2|^2 + |a_3|^2 + \cdots$ converges. The inner product of (a_1, a_2, a_3, \dots) and (b_1, b_2, b_3, \dots) is $a_1 b_1 + a_2 b_2 + a_3 b_3 + \cdots$ (which can be shown to converge by the CAUCHY-SCHWARZ INEQUALITY [V.22].)
- (ii) $L_2[0, 2\pi]$ is the set of all functions f defined on the interval $[0, 2\pi]$ of all real numbers between 0 and 2π , such that the integral $\int_0^{2\pi} |f(x)|^2 dx$ makes sense and is finite. The inner product of two functions f and g is defined to be $\int_0^{2\pi} f(x)g(x) dx$. (For technical reasons, this definition is not quite accurate, as a nonzero function can have norm zero, but this problem can easily be dealt with.)

The second of these examples is central to Fourier analysis. A *trigonometric function* is a function of the form $\cos(mx)$ or $\sin(nx)$. The inner product of any two different trigonometric functions is zero, so they are all orthogonal. Even more importantly, the trigonometric functions serve as a coordinate system for the space $L_2[0, 2\pi]$, in that every function f in the space can be represented as an (infinite) linear combination of trigonometric functions. This allows Hilbert spaces to model sound waves: if the function f represents a sound wave, then the trigonometric functions are the pure tones that are its constituent parts.

These properties of trigonometric functions illustrate a very important general phenomenon in the theory of Hilbert spaces: that every Hilbert space has an *orthonormal basis*. This means a set of vectors e_i with the following three properties:

- $\|e_i\| = 1$ for every i ;
- $\langle e_i, e_j \rangle = 0$ whenever $i \neq j$; and
- every vector v in the space can be expressed as a convergent sum of the form $\sum_i \lambda_i e_i$.

The trigonometric functions do not quite form an orthonormal basis of $L_2[0, 2\pi]$ but suitable multiples of them do. There are many contexts besides Fourier analysis where one can obtain useful information about

PUP note: this paragraph added since proofreading proof was sent.

a vector by decomposing it in terms of a given orthonormal basis, and many general facts that can be deduced from the existence of such bases.

Hilbert spaces (with complex scalars) are also central to quantum mechanics. The vectors of a Hilbert space can be used to represent possible states of a quantum mechanical system, and observable features of that system correspond to certain linear maps.

For this and other reasons, the study of LINEAR OPERATORS [III.52] on Hilbert spaces is a major branch of mathematics: see OPERATOR ALGEBRAS [IV.19]

III.38 Holomorphic Functions

A function f defined on some region D of the complex plane is called *holomorphic* if it is differentiable. This has the meaning one would expect: for every z in D the quantity $(f(z+w) - f(z))/w$ should tend to a limit as w tends to 0. This limit is denoted by $f'(z)$, and the function f' is called the *derivative* of f .

However, this bare definition hides the fact that complex differentiability is very different from real differentiability, roughly speaking because the linear approximations it gives are all of a special kind, namely “multiply by the complex number λ .” This has the effect of making complex differentiability a far stronger property than the differentiability of functions defined on \mathbb{R} or \mathbb{R}^2 . For example, if f is holomorphic, then f' is automatically holomorphic as well: the analogue of this statement for real functions is very definitely false.

Holomorphic functions are discussed in more detail in SOME FUNDAMENTAL MATHEMATICAL DEFINITIONS [I.3 §5.6].

III.39 Homology and Cohomology

If X is a TOPOLOGICAL SPACE [III.92], then one can associate with it a sequence of groups $H_n(X, R)$, where R is a commutative RING [III.83 §1] such as \mathbb{Z} or \mathbb{C} . These groups, the *homology groups* of X (with coefficients in R), are a powerful invariant: powerful because they contain a great deal of information about X but are nevertheless easy to compute, at least compared with some other invariants. The closely related *cohomology groups* $H^n(X, R)$ are more useful still because they can be made into a ring: to oversimplify slightly, an element of the cohomology group $H^n(X)$ is an EQUIVALENCE CLASS [I.2 §2.3] $[Y]$ of a subspace Y of codimension n . (Of course, for this to make true sense X should be a fairly nice space such as a MANIFOLD [I.3 §6.9].)

Then, if $[Y]$ and $[Z]$ belong to $H^n(X, R)$ and $H^m(X, R)$, respectively, their product is $[Y \cap Z]$. Since $Y \cap Z$ “typically” has codimension $n + m$, the equivalence class $[Y \cap Z]$ belongs to $H^{n+m}(X, R)$. Homology and cohomology groups are described in more detail in ALGEBRAIC TOPOLOGY [IV.10].

The concepts of homology and cohomology have become far more general than the above discussion suggests, and are no longer tied to topological spaces: for instance, the notion of group cohomology is of great importance in algebra. Even within topology, there are many different homology and cohomology theories. In 1945, Eilenberg and Steenrod devised a small number of axioms that greatly clarified the area: a homology theory is any association of groups with topological spaces that satisfies these axioms, and the fundamental properties of homology theories follow from the axioms.

III.40 Homotopy Groups

If X is a TOPOLOGICAL SPACE [III.92], then a *loop* in X is a path that begins and ends at the same point; or, more formally, a continuous function $f : [0, 1] \rightarrow X$ such that $f(0) = f(1)$. The point where the path begins and ends is called the *base point*. If two loops have the same base point, they are called *homotopic* if one can be continuously deformed to the other, with all the intermediate paths living in X and beginning and ending at the given base point. For example, if X is the plane \mathbb{R}^2 , then any two paths that begin and end at $(0, 0)$ are homotopic, whereas if X is the plane with the origin removed, then whether or not two paths (that begin and end at some other point) are homotopic depends on whether or not they go around the origin the same number of times.

Homotopy is an EQUIVALENCE RELATION [I.2 §2.3], and the equivalence classes of paths with base point x form the *fundamental group* of X , relative to x , which is denoted by $\pi_1(X, x)$. If X is connected, then this does not depend on x and we can write $\pi_1(X)$ instead. The group operation is “concatenation”: given two paths that begin and end at x , their “product” is the combined path that goes along one and then the other, and the product of equivalence classes is then defined to be the equivalence class of the product. This group is a very important invariant (see for instance GEOMETRIC AND COMBINATORIAL GROUP THEORY [IV.11 §7]); it is the first in a sequence of higher-dimensional homotopy groups, which are described in ALGEBRAIC TOPOLOGY [IV.10 §§2, 3].

III.41 The Hyperbolic Plane

The *parallel postulate* of EUCLID [VI.2] states that for any straight line L in the plane and any point x not on L there is exactly one straight line M that passes through x and does not meet L . For over 2000 years a central problem in mathematics was to decide whether this statement could be deduced from the other axioms of Euclidean geometry. Eventually, GAUSS [VI.26], BOLYAI [VI.34], and LOBACHEVSKII [VI.31] developed *hyperbolic geometry*, in which all the other axioms hold, but the parallel postulate is false because there can be more than one line through x that does not meet L . The history of this discovery is explained in GEOMETRY [II.2].

The *hyperbolic plane* can be defined in several ways. Two of the most popular are called the *half-plane model* and the *disk model*, which are RIEMANNIAN METRICS [I.3 §6.10] defined on the upper half-plane and the unit disc, respectively. Almost all the familiar concepts of Euclidean geometry can be defined for hyperbolic geometry, but their properties are different. For example, the angles of a hyperbolic triangle always add up to *less* than π . More details about the hyperbolic plane and how it is constructed can be found in SOME FUNDAMENTAL MATHEMATICAL DEFINITIONS [I.3 §§6.6, 6.10].

III.42 The Ideal Class Group

THE FUNDAMENTAL THEOREM OF ARITHMETIC [V.16] asserts that every positive integer can be written in exactly one way (apart from reordering) as a product of primes. Analogous theorems are true in other contexts as well: for example, there is a unique factorization theorem for polynomials, and another one for *Gaussian integers*, that is, numbers of the form $a + ib$ where a and b are integers.

However, for most NUMBER FIELDS [III.65], the associated “ring of integers” does not have the unique-factorization property. For example, in the RING [III.83 §1] of numbers of the form $a + b\sqrt{-5}$ with a and b integers, one can factorize 6 either as 2×3 or as $(1 + \sqrt{-5})(1 - \sqrt{-5})$.

The ideal class group is a way of measuring how badly unique factorization fails. Given any ring of integers of a number field, one can define a multiplicative structure on its set of IDEALS [III.83 §2], for which unique factorization holds. The elements of the ring itself corre-

spond to so-called “principal ideals,” so if every ideal is principal, then unique factorization holds for the ring. If there are nonprincipal ideals, then one can define a natural EQUIVALENCE RELATION [I.2 §2.3] on them in such a way that the equivalence classes, which are called *ideal classes*, form a GROUP [I.3 §2.1]. This group is the ideal class group. All principal ideals belong to the class that forms the identity of this group, so the larger and more complex the ideal group is, the further the ring is from having the unique-factorization property. For more details, see ALGEBRAIC NUMBERS [IV.3], and in particular section 7.

III.43 Irrational and Transcendental Numbers

Ben Green

An irrational number is one that cannot be written as a/b with both a and b integers. A great many naturally occurring numbers, such as $\sqrt{2}$, e , and π , are irrational. The following proof that $\sqrt{2}$ is irrational is one of the best-known arguments in all of mathematics. Suppose that $\sqrt{2} = a/b$; since common factors can be canceled, we may assume that a and b have no common factor; we have $a^2 = 2b^2$, which means that a must be even; write $a = 2c$; but then $4c^2 = 2b^2$, which implies that $2c^2 = b^2$, and hence b must be even too; this, however, is contrary to our assumption that a and b were coprime.

Several famous conjectures in mathematics ask whether certain specific numbers are rational or not. For example, $\pi + e$ and π^e are not known to be irrational, and neither is Euler’s constant:

$$\gamma = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{2} + \cdots + \frac{1}{n} \right) - \log n \approx 0.577215 \dots$$

It is known that $\zeta(3) = 1 + 2^{-3} + 3^{-3} + \cdots$ is irrational. Almost certainly, $\zeta(5), \zeta(7), \zeta(9), \dots$ are all irrational as well. However, although it has been shown that infinitely many of these numbers are irrational, no specific one is known to be.

A classic proof is that of the irrationality of e . If

$$e = \sum_{j=0}^{\infty} \frac{1}{j!}$$

were equal to p/q , then we would have

$$p(q-1)! = \sum_{j=0}^{\infty} \frac{q!}{j!}.$$

The left-hand side and the terms of the sum with $j \leq q$ are all integers. Therefore the quantity

$$\sum_{j \geq q+1} \frac{q!}{j!} = \frac{1}{q+1} + \frac{1}{(q+1)(q+2)} + \dots$$

is also an integer. But it is not hard to show that this quantity lies strictly between 0 and 1, a contradiction.

The principle used here, that a nonzero integer must have absolute value at least one, is surprisingly powerful in the theory of irrational and transcendental numbers.

Some numbers are more irrational than others. In a sense, the most irrational number is $\tau = \frac{1}{2}(1 + \sqrt{5})$, the golden ratio, because the best rational approximations to it, which are ratios of consecutive Fibonacci numbers, approach it rather slowly. There is also a very elegant proof that τ is irrational. This is based on the observation that the $\tau \times 1$ rectangle R may be divided into a square of side 1 and a $1/\tau \times 1$ rectangle. If τ were rational, then we would be able to create a rectangle with integer sides that was similar to R . From this we could remove a square, and we would be left with a smaller rectangle with integer sides that would still be similar to R . We could continue this process ad infinitum, which is clearly impossible.

A *transcendental* number is one which is not *algebraic*, that is to say, is not the root of a polynomial equation with integer coefficients. Thus $\sqrt{2}$ is not transcendental, since it solves $x^2 - 2 = 0$, and neither is $\sqrt{7 + \sqrt{17}}$.

Are there, in fact, any transcendental numbers? This question was answered by LIOUVILLE [VI.39] in 1844, who showed that various numbers were transcendental, of which

$$\begin{aligned} \kappa &= \sum_{n \geq 1} 10^{-n!} \\ &= 0.1100010000000000000000010\dots \end{aligned}$$

is a well-known example. This is not algebraic, because it can be approximated more accurately by rationals than any algebraic number can. For example, the rational approximation $110\,001/1\,000\,000$ is very close indeed to κ , but its denominator is not particularly large.

Liouville showed that if α is a root of a polynomial of degree n , then

$$\left| \alpha - \frac{a}{q} \right| > \frac{C}{q^n}$$

for all integers a and q and for some constant C depending on α . In words, α cannot be too well approximated by rationals. Roth later proved that the exponent

n here can actually be replaced by $2 + \varepsilon$ for any $\varepsilon > 0$. (For more on these topics, see LIOUVILLE'S THEOREM AND ROTH'S THEOREM [V.25].)

A completely different approach to the existence of transcendental numbers was discovered by CANTOR [VI.54] thirty years later. He proved that the set of algebraic numbers is COUNTABLE [III.11], which means, roughly speaking, that they may be listed in order. More precisely, there is a surjective map from \mathbb{N} , the set of natural numbers, to the set of algebraic numbers.

By contrast, the real numbers \mathbb{R} are not countable. Cantor's famous proof of this uses a diagonalization argument to show that any listing of all the real numbers must be incomplete.

There must, therefore, be real numbers that are not algebraic.

It is generally rather difficult to prove that a specific number is transcendental. For instance, it is by no means the case that all transcendental numbers are very well approximated by rationals; this merely provides a useful sufficient condition. There are other ways to establish that numbers are transcendental. Both e and π are known to be transcendental, and it is known that $|e - a/b| > C(\varepsilon)/b^{2+\varepsilon}$ for all $\varepsilon > 0$, so e is not all that well approximated by rationals. Since $\zeta(2m)$ is always a rational multiple of π^{2m} , it follows that the numbers $\zeta(2), \zeta(4), \dots$ are all transcendental.

The modern theory of transcendental numbers contains a wealth of beautiful results. An early one is the Gel'fond-Schneider theorem, which says that α^β is transcendental if $\alpha \neq 0, 1$ is algebraic, and if β is algebraic but not rational. In particular, $\sqrt{2}^{\sqrt{2}}$ is transcendental. There is also the *six-exponentials theorem*, which states that if x_1, x_2 are two linearly independent complex numbers, and if y_1, y_2, y_3 are three linearly independent complex numbers, then at least one of the six numbers

$$e^{x_1 y_1}, e^{x_1 y_2}, e^{x_1 y_3}, e^{x_2 y_1}, e^{x_2 y_2}, e^{x_2 y_3}$$

is transcendental. Related to this is the (as yet unsolved) *four-exponentials conjecture*: if x_1 and x_2 are two linearly independent complex numbers, and if y_1 and y_2 are linearly independent, then at least one of the four exponentials

$$e^{x_1 y_1}, e^{x_1 y_2}, e^{x_2 y_1}, e^{x_2 y_2}$$

is transcendental.

III.44 The Ising Model

PUP note: this paragraph and the following four have been rewritten and rearranged.

PUP: (odd) apostrophe is OK here.

The Ising model is one of the fundamental models of statistical physics. It was originally designed as a model for the behavior of a ferromagnetic material when it is heated up, but it has since been used to model many other phenomena.

The following is a special case of the model. Let G_n be the set of all pairs of integers with absolute value at most n . A *configuration* is a way of assigning to each point x in G_n a number σ_x , which equals 1 or -1 . The points represent atoms and $\sigma(x)$ represents whether x has “spin up” or “spin down.” With each configuration σ we associate an “energy” $E(\sigma)$, which equals $-\sum \sigma_x \sigma_y$, where the sum is taken over all pairs of neighboring points x and y . Thus, the energy is high if many points have different signs from some of their neighbors, and low if G_n is divided into large clusters of points with the same sign.

Each configuration is assigned a probability, which is proportional to $e^{-E(\sigma)/T}$. Here, T is a positive real number that represents temperature. The probability of a given configuration is therefore higher when it has small energy, so there is a tendency for a typical configuration to have clusters of points with the same sign. However, as the temperature T increases, this clustering effect becomes smaller since the probabilities become more equal.

The *two-dimensional Ising model with zero potential* is the limit of this model as n tends to infinity. For a more detailed discussion of the general model and of the *phase transition* associated with it, see PROBABILISTIC MODELS OF CRITICAL PHENOMENA [IV.26 §5].

III.45 Jordan Normal Form

Suppose that you are presented with an $n \times n$ real or complex MATRIX [I.3 §4.2] A and would like to understand it. You might ask how it behaves as a LINEAR MAP [I.3 §4.2] on \mathbb{R}^n or \mathbb{C}^n , or you might wish to know what the powers of A are. In general, answering these questions is not particularly easy, but for some matrices it is very easy. For example, if A is a *diagonal* matrix (that is, one whose nonzero entries all lie on the diagonal), then both questions can be answered immediately: if x is a vector in \mathbb{R}^n or \mathbb{C}^n , then Ax will be the vector obtained by multiplying each entry of x by the corresponding diagonal element of A , and to compute A^m you just raise each diagonal entry to the power m .

So, given a linear map T (from \mathbb{R}^n to \mathbb{R}^n or from \mathbb{C}^n to \mathbb{C}^n), it is very nice if we can find a basis with respect to which T has a diagonal matrix; if this can be done,

then we feel that we “understand” the linear map. Saying that such a basis exists is the same as saying there is a basis consisting of EIGENVECTORS [I.3 §4.3]: a linear map is called *diagonalizable* if it has such a basis. Of course, we may apply the same terminology to a matrix (since a matrix A determines a linear map on \mathbb{R}^n or \mathbb{C}^n , by mapping x to Ax). So a matrix is also called diagonalizable if it has a basis of eigenvectors, or equivalently if there is an invertible matrix P such that $P^{-1}AP$ is diagonal.

Is every matrix diagonalizable? Over the reals, the answer is no for uninteresting reasons, since there need not even be any eigenvectors: for example, a rotation in the plane clearly has no eigenvectors. So let us restrict our attention to matrices and linear maps over the complex numbers.

If we have a matrix A , then its *characteristic polynomial*, namely $\det(A - tI)$, certainly has a root, by THE FUNDAMENTAL THEOREM OF ALGEBRA [V.15]. If λ is such a root, then standard facts from linear algebra tell us that $A - \lambda I$ is singular, and therefore that there is a vector x such that $(A - \lambda I)x = 0$, or equivalently that $Ax = \lambda x$. So we do have at least one eigenvector. Unfortunately, however, there need not be enough eigenvectors to form a basis. For example, consider the linear map T that sends $(1, 0)$ to $(0, 1)$ and $(0, 1)$ to $(0, 0)$. The matrix of this map (with respect to the obvious basis) is $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. This matrix is not diagonalizable. One way of seeing why not is the following. The characteristic polynomial turns out to be t^2 , of which the only root is 0. An easy computation reveals that if $Ax = 0$ then x has to be a multiple of $(0, 1)$, so we cannot find two linearly independent eigenvectors. A rather more elegant method of proof is to observe that T^2 is the zero matrix (since it maps each of $(1, 0)$ and $(0, 1)$ to $(0, 0)$), so that if T were diagonalizable, then its diagonal matrix would have to be zero (since any nonzero diagonal matrix has a nonzero square), and therefore T would have to be the zero matrix, which it is not.

The same argument shows that *any* matrix A such that $A^k = 0$ for some k (such matrices are called *nilpotent*) must fail to be diagonalizable, unless A is itself the zero matrix. This applies, for example, to any matrix that has all of its nonzero entries below the main diagonal.

What, then, *can* we say about our nondiagonalizable matrix T above? In a sense, one feels that $(1, 0)$ is “nearly” an eigenvector, since we do have $T^2(1, 0) = (0, 0)$. So what happens if we extend our point of view

by allowing such vectors? One would say that a vector x is a *generalized eigenvector* of T , with eigenvalue λ , if some power of $T - \lambda$ maps x to zero. For instance, in our example above the vector $(1, 0)$ is a generalized eigenvector with eigenvalue 0. And, just as we have an “eigenspace” associated with each eigenvalue λ (defined to be the space of all eigenvectors with eigenvalue λ), we also have a “generalized eigenspace,” which consists of all generalized eigenvectors with eigenvalue λ .

Diagonalizing a matrix corresponds exactly to decomposing the vector space (\mathbb{C}^n) into eigenspaces. So it is natural to hope that one could decompose the vector space into generalized eigenspaces for *any* matrix. And this turns out to be true. The way of breaking up the space is called *Jordan normal form*, which we shall now describe in more detail.

Let us pause for a moment and ask: what is the very simplest situation in which we get a generalized eigenvector? It would surely be the obvious generalization of the above example to n dimensions. In other words, we have a linear map T that sends e_1 to e_2 , e_2 to e_3 , and so on, until e_{n-1} is sent to e_n , with e_n itself mapped to zero. This corresponds to the matrix

$$\begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix}.$$

Although this matrix is not diagonalizable, its behavior is at least very easy to understand.

The Jordan normal form of a matrix will be a diagonal sum of matrices that are easily understood in the way that this one is. Of course, we have to consider eigenvalues other than zero: accordingly, we define a *block* to be any matrix of the form

$$\begin{pmatrix} \lambda & 0 & 0 & \cdots & 0 & 0 \\ 1 & \lambda & 0 & \cdots & 0 & 0 \\ 0 & 1 & \lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & \lambda \end{pmatrix}.$$

Note that this matrix A , with λI subtracted, is precisely the matrix above, so that $(A - \lambda I)^n$ is indeed zero. Thus, a block represents a linear map that is indeed easy to understand, and all its vectors are generalized eigenvectors with the same eigenvalue. The Jordan normal form theorem tells us that every matrix can be decomposed into such blocks: that is, a matrix is in Jordan

normal form if it is of the form

$$\begin{pmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_k \end{pmatrix}.$$

Here, the B_i are blocks, which can have different sizes, and the 0s represent submatrices of the matrix with sizes depending on the block sizes. Note that a block of size 1 simply consists of an eigenvector.

Once a matrix A is put into Jordan normal form, we have broken up the space into subspaces on which it is easy to understand the action of A . For example, suppose that A is the matrix

$$\begin{pmatrix} 4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 2 \end{pmatrix},$$

which is made out of three blocks, of sizes 3, 2, and 2. Then we can instantly read off a great deal of information about A . For instance, consider the eigenvalue 4. Its algebraic multiplicity (its multiplicity as a root of the characteristic polynomial) is 5, since it is the sum of the sizes of all the blocks with eigenvalue 4, while its geometric multiplicity (the dimension of its eigenspace) is 2, since it is the *number* of such blocks (because in each block we only have one actual eigenvector). And even the minimum polynomial of the matrix (the smallest-degree polynomial $P(t)$ such that $P(A) = 0$) is easy to write down. The minimum polynomial of each block can be written down instantly: if the block has size k and generalized eigenvalue λ , then it is $(t - \lambda)^k$. The minimum polynomial of the whole matrix is then the “lowest common multiple” of the polynomials for the individual blocks. For the matrix above, we get $(t - 4)^3$, $(t - 4)^2$, and $(t - 2)^2$ for the three blocks, so the minimum polynomial of the whole matrix is $(t - 4)^3(t - 2)^2$.

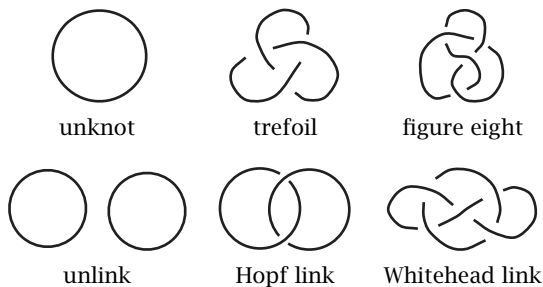
There are some generalizations of Jordan normal form, away from the context of linear maps acting on vector spaces. For example, there is an analogue of the theorem that applies to Abelian groups, which turns out to be the statement that every finite Abelian group can be decomposed as a direct product of cyclic groups.

III.46 Knot Polynomials

W. B. R. Lickorish

1 Knots and Links

A *knot* is a curve in three-dimensional space that is closed (in other words, it stops where it began) and never meets itself along its way. A *link* is several such curves, all disjoint from one another, which are called the *components* of the link. Some simple examples of knots and links are the following:



Two knots are equivalent or “the same” if one can be moved continuously, never breaking the “string,” to become the other. *Isotopy* is the technical term for such movement. For example, the following knots are the same:



The first problem in knot theory is how to decide if two knots are the same. Two knots may appear to be very different but how does one *prove* that they are different? In classical geometry two triangles are the same (or *congruent*) if one can be moved rigidly on to the other. Numbers that measure side-lengths and angles are assigned to each triangle to help determine if this is the case. Similarly, mathematical entities called *invariants* can be associated with knots and links in such a way that if two links have different invariants, then they cannot be the same link. Many invariants relate to the geometry or topology of the complement of a link in three-dimensional space. The FUNDAMENTAL GROUP [IV.10 §2] of this complement is an excellent invariant, but algebraic techniques are then needed to distinguish the groups. The polynomial of J. W. Alexander (published in 1926) is a link invariant derived from distinguishing such groups. Although rooted in ALGEBRAIC TOPOLOGY [IV.10], the Alexander

polynomial has long been known to satisfy a skein relation (see below). The HOMFLY polynomial of 1984 generalizes the Alexander polynomial and can be based on the simple combinatorics of skein theory alone.

1.1 The HOMFLY Polynomial

Suppose that links are oriented so that directions, indicated by arrows, are given to all components. To each oriented link L is assigned its HOMFLY polynomial $P(L)$, a polynomial with integer coefficients in two variables v and z (allowing both positive and negative powers of v and z). The polynomials are such that

$$P(\text{unknot}) = 1 \quad (1)$$

and there is a linear *skein relation*

$$v^{-1}P(L_+) - vP(L_-) = zP(L_0). \quad (2)$$

This means that whenever three links have identical diagrams except near one crossing, where they are as follows



then this equation holds.

This turns out to be good notation, although one could in principle use x and y in place of v^{-1} and $-v$. Although Alexander's polynomial satisfied a particular instance of (2), it took almost sixty years and the discovery of the Jones polynomial for it to be realized that this general linear relation can be used. Note that there are two possible types of crossing in a diagram of an oriented link. A crossing is *positive* if, when approaching the crossing along the under-passing arc in the direction of the arrow, the other directed arc is seen to cross over from left to right. If the over-passing arc crosses from right to left, the crossing is *negative*. When interpreting the skein relation at a crossing of a link L , it is vital that L be regarded as L_+ if the crossing is positive and as L_- if it is negative.

The theorem that underpins this theory, which is not at all obvious, is that it is possible to assign such polynomials to oriented links in a coherent fashion, uniquely, independent of any choice of a link's diagram. A proof of this is given in Lickorish (1997).

1.2 HOMFLY Calculations

In a diagram of a knot it is always possible to change some of the crossings, from over to under, to achieve a diagram of the unknot. Links can be undone similarly. Using this, the polynomial of any link can be calculated

from the above equations, though the length of the calculation is exponential in the number of crossings. The following is a calculation of $P(\text{trefoil})$. Firstly, consider the following instance of the skein relation:

$$v^{-1}P(\text{trefoil}) - vP(\text{trefoil}) = zP(\text{unknot}).$$

Substituting the polynomial 1 for the polynomials of the two unknots, this shows that the HOMFLY polynomial of the two-component unlink is $z^{-1}(v^{-1} - v)$. A second usage of the skein relation is

$$v^{-1}P(\text{Hopf link}) - vP(\text{Hopf link}) = zP(\text{unknot}).$$

Substituting the previous answer for the unlink shows that the HOMFLY polynomial of the Hopf link is $z^{-1}(v^{-3} - v^{-1}) - zv^{-1}$. Finally, consider the following instance of the skein relation:

$$v^{-1}P(\text{trefoil}) - vP(\text{trefoil}) = zP(\text{trefoil}).$$

Substitution of the polynomial for the Hopf link already calculated and, of course, the value 1 for the unknot shows that

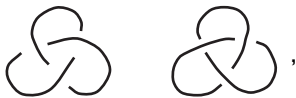
$$P(\text{trefoil}) = -v^{-4} + 2v^{-2} + z^2v^{-2}.$$

A similar calculation shows that

$$P(\text{figure eight}) = v^2 - 1 + v^{-2} - z^2.$$

The trefoil and the figure eight thus have different polynomials; this *proves* they are different knots. Experimentally, if a trefoil is actually made from a necklace (using the clasp to join the ends together) it is indeed found to be impossible to move it to the configuration of a figure eight knot. Note that the polynomial of a knot is not dependent on the choice of its orientation (but this is not so for links).

Reflecting a knot in a mirror is equivalent to changing every crossing in a diagram of the knot from an over-crossing to an under-crossing and vice versa (consider the plane of the diagram to be the mirror). The polynomial of the reflection is always the same as that of the original knot *except* that every occurrence of v must be replaced by one of $-v^{-1}$. Thus the trefoil and its reflection,



have polynomials

$$-v^{-4} + 2v^{-2} + z^2v^{-2} \quad \text{and} \quad -v^4 + 2v^2 + z^2v^2.$$

As these polynomials are not the same, the trefoil and its reflection are different knots.

2 Other Polynomial Invariants

The HOMFLY polynomial was inspired by the discovery in 1984 of the polynomial of V. F. R. Jones. For an oriented link L , the Jones polynomial $V(L)$ has just one variable t (together with t^{-1}). It is obtained from $P(L)$ by substituting $v = t$ and $z = t^{1/2} - t^{-1/2}$, where $t^{1/2}$ is just a formal square root of t . The Alexander polynomial is obtained by the substitution $v = 1$, $z = t^{-1/2} - t^{1/2}$. This latter polynomial is well understood in terms of topology, by way of the fundamental group, covering spaces, and homology theory, and can be calculated by various methods involving determinants. It was J. H. Conway who, in discussing in 1969 his normalized version of the Alexander polynomial (the polynomial in one variable z obtained by substituting $v = 1$ into the HOMFLY polynomial), first developed the theory of skein relations.

There is one more polynomial (due to L. H. Kauffman) based on a linear skein relation. The relation involves four links with unoriented diagrams differing as follows:



There are examples of pairs of knots that the Kauffman polynomial but not the HOMFLY polynomial can distinguish and vice versa; some pairs are not distinguished by any of these polynomials.

2.1 Application to Alternating Knots

For the Jones polynomial there is a particularly simple formulation, by means of "Kauffman's bracket polynomial," that leads to an easy proof that the Jones (but *not* the HOMFLY) polynomial is coherently defined. This approach has been used to give the first rigorous confirmation of P. G. Tait's (1898) highly believable proposal that a reduced alternating diagram of a knot has the minimal number of crossings for any diagram of that knot. Here "alternating" means that in going along the knot the crossings go: ... over, under, over, under, over, ... Not every knot has such a diagram. "Reduced" means that there are, adjacent to each crossing, four *distinct* regions of the diagram's planar complement. Thus, for example, any nontrivial reduced alternating diagram is not a diagram of the unknot. Also, the figure eight knot certainly has no diagram with only three crossings.

2.2 Physics

Unlike that of Alexander, the HOMFLY polynomial has no known interpretation in terms of classical algebraic topology. It can, however, be reformulated as a collection of state sums, summing over certain labelings of a knot diagram. This recalls ideas from statistical mechanics; an elementary account is given in Kauffman (1991). An amplification of the whole HOMFLY polynomial theory leads into a version of conformal field theory called topological quantum field theory.

Further Reading

- Kauffman, L. H. 1991. *Knots and Physics*. Singapore: World Scientific.
- Lickorish, W. B. R. 1997. *An Introduction to Knot Theory*. Graduate Texts in Mathematics, volume 175. New York: Springer.
- Tait, P. G. 1898. On knots. In *Scientific Papers*, volume I, pp. 273–347. Cambridge: Cambridge University Press.

III.47 K-Theory

K-theory concerns one of the most important invariants of a TOPOLOGICAL SPACE [III.92] X , a pair of groups called the *K*-groups of X . To form the group $K^0(X)$ one takes all (equivalence classes of) vector bundles on X , and uses the direct sum as the group operation. This leads not to a group but to a semigroup. However, from the semigroup one can easily construct a group in the same way that one constructs \mathbb{Z} out of \mathbb{N} : by taking equivalence classes of expressions of the form $a - b$. If i is a positive integer, then there is a natural way of defining a group $K^{-i}(X)$: it is closely related to the group $K^0(S^i \times X)$. The very important *Bott periodicity theorem* says that $K^i(X)$ depends only on the parity of i , so there are in fact just two distinct *K*-groups, $K^0(X)$ and $K^1(X)$. See ALGEBRAIC TOPOLOGY [IV.10 §6] for more details.

If X is a topological space such as a compact manifold, then one can associate with it the C^* -algebra $C(X)$ of all continuous functions from X to \mathbb{C} . It turns out to be possible to define the *K*-groups in terms of this algebra in such a way that it applies to algebras that are not of the form $C(X)$. In particular, it applies to algebras where multiplication is not commutative. For instance, *K*-theory provides important invariants of C^* -algebras. See OPERATOR ALGEBRAS [IV.19 §4.4].

Lagrange Multipliers

See OPTIMIZATION AND LAGRANGE MULTIPLIERS [III.66]

III.48 The Leech Lattice

To define a *lattice* in \mathbb{R}^d one chooses d linearly independent vectors v_1, \dots, v_d and takes all combinations of the form $a_1 v_1 + \dots + a_d v_d$, where a_1, \dots, a_d are integers. For example, to define the *hexagonal lattice* in \mathbb{R}^2 one can take v_1 and v_2 to be $(1, 0)$ and $(\frac{1}{2}, \sqrt{\frac{3}{2}})$, respectively. Notice that v_2 is v_1 rotated by $\pi/3$, and also that $v_2 - v_1$ is v_2 rotated by $\pi/3$. Continuing this process, one can generate all the points in a regular hexagon about the origin.

The hexagonal lattice is unusual, among lattices in \mathbb{R}^2 , in that it has a rotational symmetry of order 6. This makes it the “best” lattice in many ways. (For example, bees arrange their hives in hexagonal lattices, soap bubbles of similar sizes naturally organize themselves into hexagonal lattices, and so on.) The Leech lattice plays a similar role in twenty-four dimensions: it is the “most symmetrical” of all twenty-four-dimensional lattices, with a degree of symmetry that is quite extraordinary. It is discussed in more detail in THE GENERAL GOALS OF MATHEMATICAL RESEARCH [I.4 §4].

III.49 L-Functions

Kevin Buzzard

1 How Can We “Package” a Sequence of Numbers?

Suppose we are given a sequence of numbers such as

$$\pi, \sqrt{2}, 6.023 \times 10^{23}, \dots$$

How can we package up this sequence into *one* object that remembers everything about the sequence, and that might even give us new insights into the sequence? One standard technique is to use a GENERATING FUNCTION [III.32], but here is another way, which has proved very fruitful in number theory and elsewhere. Given a sequence a_1, a_2, a_3, \dots , we define the *Dirichlet series*

$$\begin{aligned} L(s) &= \frac{a_1}{1^s} + \frac{a_2}{2^s} + \frac{a_3}{3^s} + \dots \\ &= \sum_{n \geq 1} a_n / n^s. \end{aligned}$$

Here, s could be a positive integer, or a real number, for example. As long as our sequence a_1, a_2, \dots does not

grow too quickly (which we shall henceforth assume), the series $L(s)$ will converge for all sufficiently large values of s . Moreover, it may be a very “rich” object, even if the initial sequence is simple. For example, if $a_n = 1$ for all n , then the resulting function $L(s)$ is the famous RIEMANN ZETA FUNCTION [IV.4 §3] $\zeta(s) = 1^{-s} + 2^{-s} + 3^{-s} + \cdots$, which converges when $s > 1$ and was shown by Euler to satisfy the following identities, among others (there is one for each even number):

$$\begin{aligned}\zeta(2) &= \pi^2/6, & \zeta(4) &= \pi^4/90, \\ \zeta(12) &= \frac{691\pi^{12}}{638\,512\,875}.\end{aligned}$$

Thus, even a sequence as simple as $1, 1, 1, \dots$ leads us to some natural questions that cry out to be answered.

The zeta function is the prototypical example of an *L*-function. However, not every Dirichlet series deserves to be called an *L*-function. We will mention below some “good” properties that the zeta function has: roughly speaking, a Dirichlet series is considered to be an *L*-function if it has these good properties. This is not a formal definition of course, but in fact there is no formal definition of “an *L*-function.” (People have tried to give one, but there is no real consensus about what the right definition should be.) What happens in practice is that a mathematician finds a way of associating a sequence a_1, a_2, \dots of numbers with a mathematical object X , and if evidence then emerges to suggest that the associated Dirichlet series $L(s)$ shares the good properties of the zeta function, then $L(s)$ will be called the *L*-function of X .

2 What Good Properties Might $L(s)$ Have?

One can check that the zeta function can also be expressed as an infinite product over primes $\zeta(s) = \prod_p (1 - p^{-s})^{-1}$. The product is usually referred to as an *Euler product*, and if a Dirichlet series is to deserve the title of *L*-function, then it should have some kind of analogous product expansion. The existence of such an expansion is closely related to, but a little stronger than, the property that the sequence a_1, a_2, \dots should be *multiplicative*, which means that $a_{mn} = a_m a_n$ whenever m and n are coprime.

To go further we must expand our horizons. It is not hard to show that our definition of $L(s)$ makes sense even when s is a *complex* number, as long as it has a sufficiently large real part. Moreover, it defines a HOLOMORPHIC FUNCTION [I.3 §5.6] in the region of the complex plane where the sum converges. For example, the Dirichlet series defining the zeta function converges for

every s such that $\operatorname{Re}(s) > 1$. A standard fact about the zeta function is that it has a unique extension to a holomorphic function of s for *any* complex number $s \neq 1$. This phenomenon is known as *meromorphic continuation* of the zeta function. It is similar to the fact that the infinite sum $1 + x + x^2 + x^3 + \cdots$ converges only when $|x| < 1$ but, when rewritten as $1/(1 - x)$, has a natural interpretation for any complex number x other than 1. A meromorphic continuation is another of the properties that one would expect of a general *L*-function. It is important to stress, however, that extending a Dirichlet series to a function on the whole complex plane is *not* a “purely formal” technique: for a random sequence a_1, a_2, \dots there is no reason at all for the associated Dirichlet series $L(s)$ to have a natural extension beyond the region where the series converges. The existence of a meromorphic continuation is somehow a rigorous way of asserting the existence of subtle symmetries in the series.

While on the subject of meromorphic continuation, we should briefly mention THE RIEMANN HYPOTHESIS [V.33], a conjecture which states that, once one has extended $\zeta(s)$ to a function on the whole complex plane, the complex numbers s such that $0 < \operatorname{Re}(s) < 1$ and $\zeta(s) = 0$ all have real part equal to $\frac{1}{2}$. There are analogous Riemann hypotheses for many *L*-functions, almost all of which are open problems.

The final property we shall emphasize is that there is a relatively simple formula relating $\zeta(s)$ and $\zeta(1 - s)$. This relation is called the *functional equation* of the zeta function, and any Dirichlet series worthy of the name *L*-function should also have an analogous property. (In general one looks for a relation between $L(s)$ and $\bar{L}(k - s)$, where k is some real number and $\bar{L}(s)$ is the Dirichlet series associated with the series of complex conjugates $\overline{a_1}, \overline{a_2}, \dots$)

There are many examples of Dirichlet series arising in number theory that do have, or are at least conjectured to have, these three key properties: an Euler product, meromorphic continuation, and a functional equation. These are the Dirichlet series that have come to be known as *L*-functions. For example, if A and B are integers such that the three roots of the cubic polynomial $x^3 + Ax + B$ are distinct, then the equation

$$y^2 = x^3 + Ax + B \tag{1}$$

defines an ELLIPTIC CURVE [III.21], and there is a natural sequence a_1, a_2, \dots associated with it (a_n is related to the number of solutions of (1) modulo n , at least when n is prime—see ARITHMETIC GEOMETRY

[IV.6 §5.1] for more details). However, it was an open problem for years to establish the existence of a meromorphic continuation of the associated Dirichlet series $L(s)$ to the complex plane: it is now known to exist (and indeed to have no poles) as a consequence of the work of Wiles, Taylor, and others that grew out of the proof of FERMAT'S LAST THEOREM [V.12].

3 What Is the Point of L -Functions?

One of the first uses of L -functions was by DIRICHLET [VI.36] himself, who used them to prove that there are infinitely many primes in a general arithmetic progression (see ANALYTIC NUMBER THEORY [IV.4 §4]). In fact, although the Riemann hypothesis is still an open problem, even partial results about the locations of the zeros of the Riemann zeta function have deep consequences in the theory of distribution of prime numbers.

However, over the last hundred years mathematicians have realized a second use for them: if X is a mathematical object and $L(s)$ is its associated L -function, then there are deep conjectures relating the arithmetic of X to the values that $L(s)$ assumes, typically at points where the Dirichlet series defining $L(s)$ does not converge! Hence, one can investigate X by investigating its L -function. One basic example of this phenomenon is the BIRCH-SWINNERTON-DYER CONJECTURE [V.4], a weak form of which states that the L -function associated with equation (1) should vanish at $s = 1$ if and only if (1) has infinitely many solutions such that both x and y are rational numbers. Much is known about this conjecture, and it has been vastly generalized by work of Deligne, Belinson, Bloch, and Kato. However, at the time of writing it remains open.

III.50 Lie Theory

Mark Ronan

1 Lie Groups

Why are groups important in mathematics? One major reason is that it is often possible to understand a mathematical structure by understanding its symmetries, and the symmetries of a given mathematical structure form a group. Some mathematical structures are so symmetrical that they have not just a finite number of symmetries, but a continuous family of them. When this is the case, we find ourselves in the realms of Lie groups and Lie theory.

One of the simplest “continuous” groups is the group $SO(2)$, which consists of all rotations of the plane \mathbb{R}^2 about the origin. With each element of $SO(2)$ one can associate an angle θ : the angle of the rotation in question. If we write R_θ for the counterclockwise rotation by θ , then the group operation is given by $R_\theta R_\varphi = R_{\theta+\varphi}$, where $R_{2\pi}$ is understood to equal R_0 , the identity element of the group.

The group $SO(2)$ is not just a continuous group, but also a *Lie group*. Roughly speaking, this means that it is a group in which one can meaningfully define the concept of a smooth curve (that is, a curve that is not just continuous but differentiable as well). Given any two elements R_θ and R_φ of $SO(2)$, one can easily define a smooth path from R_θ to R_φ by smoothly modifying θ until it becomes φ . (The most obvious such path would be given in parametric form by $R_{(1-t)\theta+t\varphi}$, as t goes from 0 to 1.) It is not always the case that every pair of points in a Lie group can be connected by a path: when they can, the Lie group is said to be *connected*. An example of a Lie group that is not connected is $O(2)$, which consists of $SO(2)$ together with all reflections of the plane about lines through the origin. Any two rotations can be linked by a path, as can any two reflections, but there is no continuous way of changing a rotation into a reflection.

Lie groups were introduced by SOPHUS LIE [VI.53] in order to create an analogue of GALOIS THEORY [V.24] for differential equations. Lie groups that consist of invertible linear transformations of \mathbb{R}^n or \mathbb{C}^n , like the examples above, are called *linear Lie groups*, and they are an important subclass. For linear Lie groups it is fairly easy to work out what terms such as “continuous,” “differentiable,” or “smooth” should mean. However, one can also consider more abstract Lie groups (both real and complex), with elements that are not given as linear transformations. In order to give a proper definition of Lie groups in their full generality, one needs the concept of a smooth MANIFOLD [I.3 §6.9]. However, for simplicity we shall mostly restrict attention to linear Lie groups.

A very common way to create a Lie group is to collect all transformations of a given space that preserve one or more specified geometric structures. For instance, the *general linear group* $GL_n(\mathbb{R})$ is defined to be the group of all invertible linear transformations from \mathbb{R}^n to \mathbb{R}^n . Inside this group is the *special linear group* $SL_n(\mathbb{R})$, in which we retain only those linear transformations that preserve volume and orientation (or equivalently those with DETERMINANT [III.15] equal

PUP: this as you wanted it? Or 'Sophus LIE'? Part VI not alphabetical so your note about this appearing 'under Lie' is not really relevant. (T&T note: if change needed, see also Cartan CR later in this article and search for 'SIMON STEVIN [VI.10]' and 'WILLIAM ROWAN' as well.)

to 1). If instead we retain the linear transformations that preserve distance, then we obtain the *orthogonal group* $O(n)$; if we retain linear transformations that preserve both distance and orientation we obtain the *special orthogonal group* $SO(n)$, which is easily seen to equal $SL_n(\mathbb{R}) \cap O(n)$. The *Euclidean group* $E(n)$ of rigid motions of \mathbb{R}^n (that is, all transformations that preserve distances and angles, such as rotations, reflections, and translations) is generated by the orthogonal group $O(n)$, together with the group of translations (which is isomorphic to \mathbb{R}^n). There are analogues of all of the above groups in which the real numbers \mathbb{R} are replaced by the complex numbers \mathbb{C} . For instance, $GL_n(\mathbb{C})$ is the group of all invertible complex-linear transformations of \mathbb{C}^n , and the complex analogue of the orthogonal group $O(n)$ is the *unitary group* $U(n)$. There are also the *symplectic groups* $Sp(2n)$, which are analogues of $O(n)$ and $U(n)$ over the QUATERNIONS [III.78]. These are all manifestly linear Lie groups except for $E(n)$, and in fact it is not difficult to describe a linear Lie group that is isomorphic to $E(n)$ as well.

Many important examples of Lie groups are finite dimensional, which roughly means that they can be described using a finite number of continuous parameters. (Infinite-dimensional Lie groups, while important, are more difficult to handle and will not be discussed in detail here.) For example, the group $SO(3)$, of rotations of \mathbb{R}^3 that fix the origin, is three dimensional. Each rotation can be specified using three parameters, which could, for instance, be taken as rotations around the x -axis, y -axis, and z -axis. These particular parameters are known to airline pilots as roll, pitch, and yaw, where the x -axis is in the direction of the airplane. Another way of specifying each rotation is by its axis and angle of rotation. Two parameters are needed to specify the axis (using spherical coordinates for example), and one parameter is needed to specify the angle of rotation. Let us take this angle to be between 0 and π (a rotation by an angle greater than π has the same effect as a rotation by an angle less than π from the opposite direction).

We can represent $SO(3)$ geometrically as follows. Let B be a ball of radius π centered at the origin. Given any noncentral point P of B , associate with it the rotation of \mathbb{R}^3 about the axis OP (where O is the origin) through an angle that is given in radians by the distance from O to P . With O itself we associate the identity map, so the only ambiguity is that a rotation through π radians is associated with two opposite points P and P' on the surface of B . We can remove this ambiguity by glu-

ing all such pairs of points together. This tells us what $SO(3)$ looks like as a TOPOLOGICAL SPACE [III.92]: it is equivalent to the three-dimensional PROJECTIVE SPACE [I.3 §6.7] \mathbb{RP}^3 . The group $SO(2)$, by comparison, is much simpler, and is topologically equivalent to a circle.

Lie groups arise naturally in any subject that involves continuous motion. For instance, they appear in applied topics such as the design of guidance systems and also in very pure topics such as geometry or differential equations. Lie groups, and the closely related Lie algebras discussed below, also frequently arise in many types of algebra, particularly in the algebraic structures that appear in quantum mechanics and other related branches of physics.

2 Lie Algebras

As the examples above show, Lie groups are often “curved” and have some nontrivial topology. However, one can profitably analyze a Lie group by associating with it a flat space known as a *Lie algebra*. This idea is similar to the idea of studying a symmetric object such as a sphere by first studying its relationship to one of its tangent planes. The Lie algebra uses the tangent space to the Lie group at the identity element, and one can view it as a “logarithm” of the Lie group.

To see how Lie algebras arise, let us consider a linear Lie group. The elements of the group can be viewed as linear transformations on a vector space, or equivalently (when we have selected a coordinate basis) as square matrices. In general, two matrices A and B do not commute (that is, AB does not have to equal BA), but the situation becomes simpler if one looks at matrices that are very close to the identity matrix I . If $A = I + \epsilon X$ and $B = I + \epsilon Y$ for some very small positive ϵ and two fixed matrices X and Y , then

$$AB = I + \epsilon(X + Y) + \epsilon^2 XY$$

and

$$BA = I + \epsilon(X + Y) + \epsilon^2 YX.$$

Thus, if we ignore the terms containing ϵ^2 , we see that A and B “almost commute,” and that multiplication of A and B “almost corresponds to” addition of X and Y : indeed, one can view X and Y as analogous to “logarithms” of A and B respectively.

Let us now informally define the *Lie algebra* \mathfrak{g} of a linear Lie group G to be the space of all matrices X such that, for sufficiently small ϵ , the matrix $I + \epsilon X$ lies in G , up to errors of size ϵ^2 . For example, the Lie algebra $\mathfrak{gl}_n(\mathbb{C})$ of the general linear group $GL_n(\mathbb{C})$ is the space of

all $n \times n$ complex matrices. One can view the Lie algebra as describing all possible instantaneous directions and speeds within the group G , and a more precise definition is the collection of all derivatives R'_0 of smooth curves $\epsilon \mapsto R_\epsilon$ in G that pass through the identity element R_0 . This definition can also be extended to more abstract Lie groups without much difficulty. (To return to the example of the airplane pilot, an element of the Lie group $SO(3)$ could be used to describe the current orientation of the aircraft with respect to a fixed coordinate system, whereas an element of the Lie algebra $\mathfrak{so}(3)$ could be used to describe the current rate of roll, pitch, and yaw that the pilot is applying to the aircraft to smoothly change its orientation.)

As we have just seen, the Lie algebra $\mathfrak{gl}_n(\mathbb{C})$ of the general linear group $GL_n(\mathbb{C})$ is the space of all $n \times n$ complex matrices. The Lie algebra $\mathfrak{sl}_n(\mathbb{C})$ of the special linear group $SL_n(\mathbb{C})$ is the subspace of all matrices with trace zero. This is because $\det(I + \epsilon X) = 1 + \epsilon \operatorname{tr} X$, up to errors of size ϵ^2 , so if $\epsilon \mapsto I + \epsilon X$ is a path in the group, then $\operatorname{tr} X = 0$. The Lie algebra $\mathfrak{so}(n)$ of $SO(n)$ is equal to the Lie algebra $\mathfrak{o}(n)$ of $O(n)$, and both are equal to the space of all antisymmetric matrices. Similarly, both the Lie algebra $\mathfrak{su}(n)$ of $SU(n)$ and the Lie algebra $\mathfrak{u}(n)$ of $U(n)$ are equal to the space of skew-Hermitian matrices. (A matrix is skew-Hermitian if it equals minus the complex conjugate of its transpose.)

The fact that a Lie group is closed under multiplication can be used to show that its Lie algebra is closed under addition. Thus, a Lie algebra is a (real) vector space. However, it has some additional structure that makes it far more than *just* a vector space. For instance, let A and B be two elements of the Lie group G that are very close to the identity. Then we can write $A \approx I + \epsilon X$ and $B \approx I + \epsilon Y$ for some very small ϵ and some elements X and Y of the Lie algebra \mathfrak{g} . A little matrix algebra shows that the *commutator* $ABA^{-1}B^{-1}$ of A and B , which is the element of G that measures the extent to which A and B fail to commute, can be approximated by $I + \epsilon^2[X, Y]$, where $[X, Y] = XY - YX$. This quantity $[X, Y]$ is called the *Lie bracket* of X and Y . Informally, it represents the net direction of motion if one first moves an infinitesimal amount in the X direction, then in the Y direction, then back in the X direction and back in the Y direction, in that order. The resulting new direction may be quite different from the original directions X and Y .

The Lie bracket obeys a number of nice identities, such as the antisymmetric identity $[X, Y] = -[Y, X]$

and the *Jacobi identity*

$$[[X, Y], Z] + [[Y, Z], X] + [[Z, X], Y] = 0.$$

One can in fact use such identities to define Lie algebras in a completely abstract fashion, without any reference to matrices or Lie groups, in much the same way that other algebraic objects such as groups, rings, and fields can be defined using a handful of algebraic identities as axioms, but we shall not focus on the abstract approach to Lie algebras here. A familiar example of a Lie algebra is \mathbb{R}^3 with the Lie bracket $[x, y]$ defined to be the cross-product $x \times y$. Notice that the Lie bracket does not satisfy the associative law (unless it is trivial).

We have seen that a linear Lie group G naturally generates the bracket operation $[\cdot, \cdot]$ on its Lie algebra \mathfrak{g} . Conversely, if the Lie group is connected, one can almost reconstruct it from the Lie algebra, with its addition, scalar multiplication, and Lie bracket operation. More precisely, every element A of the Lie group can be written as an EXPONENTIAL [III.25] $\exp(X)$ of an element X of the Lie algebra. For example, if the Lie group is $SO(2)$, then we can identify it with the unit circle in \mathbb{C} . The tangent to this circle at 1 is a vertical line, so we can identify the Lie algebra with the set $i\mathbb{R}$ of purely imaginary numbers. (Normally, however, we would just say that the Lie algebra is \mathbb{R} .) The rotation through an angle θ can then be written as $\exp(i\theta)$. Note that this representation is not unique, since $\exp(i\theta) = \exp(i(\theta + 2\pi))$. It is not hard to see that the Lie group \mathbb{R} also has \mathbb{R} as its Lie algebra (to make sense of this it helps to replace \mathbb{R} by the multiplicative group of positive real numbers, which is isomorphic to \mathbb{R}), and that in this case the representation of a group element as an exponential is unique. In general, if two connected Lie groups have the same Lie algebra, then those Lie groups share the same universal cover, and are therefore closely related to one another.

In the case of linear Lie groups, the exponential can be described by the familiar formula

$$\exp(X) = \lim_{n \rightarrow \infty} \left(I + \frac{1}{n} X \right)^n.$$

For more abstract Lie groups, the exponential is best described in the language of ordinary differential equa-

tions,¹ using a suitable generalization of the identity

$$\frac{d}{dt}e^{tX} = Xe^{tX}$$

from single-variable calculus. However, owing to the noncommutativity of the Lie group, it is not quite true that $\exp(X + Y)$ equals $\exp(X)\exp(Y)$; instead, the correct identity is the *Baker-Campbell-Hausdorff formula*

$$\exp(X)\exp(Y) = \exp(X + Y) + \frac{1}{2}[X, Y] + \cdots,$$

where the missing terms consist of a moderately complicated infinite series involving the Lie bracket. The exponential map that connects Lie algebras and Lie groups is closely related to the Lie bracket, and because of this it is possible to study and classify Lie groups by first studying and classifying Lie algebras with their Lie bracket operation.

3 Classification

It is always of interest when a mathematical structure can be classified, but especially so if the structure is important and the classification is not straightforward. By these criteria, the results that have been obtained concerning the classification of Lie algebras are undeniably interesting, and they are regarded as one of the great mathematical achievements from around the turn of the twentieth century.

It turns out to be easier to classify *complex* Lie algebras: that is, Lie algebras such as $\mathfrak{sl}_n(\mathbb{C})$ that have the structure of a complex vector space. Each real Lie algebra embeds in a complex Lie algebra of twice the (real) dimension, known as the *complexification* of the original algebra. However, a complex Lie algebra may arise as the complexification of several different real Lie algebras (known as *real forms* of the complex Lie algebra).

In classifying Lie groups and Lie algebras, the first step is to restrict attention to *simple* Lie groups and Lie algebras; these are analogous to prime numbers in the sense that they cannot be “factored” into smaller components. For instance, the Euclidean group $E(n)$ contains the translation group \mathbb{R}^n as a connected normal subgroup. If we factor out this group, then we obtain the orthogonal group $O(n)$, so $E(n)$ is not simple. More

formally, a Lie group is *simple* if it contains no proper connected normal subgroups, and a Lie algebra is *simple* if it contains no proper IDEALS [III.83 §2]. In this sense, the Lie group $SL_n(\mathbb{C})$ and its Lie algebra $\mathfrak{sl}_n(\mathbb{C})$ are simple for every n . Finite-dimensional, complex, simple Lie algebras were classified by Wilhelm Killing and Élie CARTAN [VI.69] in 1888–94.

This classification is often placed in the context of so-called *semisimple* Lie algebras, which can be factored in a unique way (up to rearrangement) as a direct sum of simple Lie algebras, just as a natural number can be factored uniquely as a product of prime numbers. Furthermore, a theorem of Levi shows that a general finite-dimensional Lie algebra \mathfrak{g} can be expressed as a combination (or, more precisely, a “semidirect product”) of a semisimple algebra (called a *Levi subalgebra* of \mathfrak{g}) and a solvable subalgebra (known as the *radical* of \mathfrak{g}). Solvable Lie algebras, which are related to the concept of a SOLVABLE GROUP [V.24] in group theory, are difficult to classify, but in many applications one can restrict attention to semisimple Lie algebras, and hence to simple Lie algebras.

A simple Lie algebra \mathfrak{g} splits into smaller subalgebras, which are not ideals but which are related to one another in particularly nice ways. The case of \mathfrak{sl}_{n+1} is typical and we shall use it to explain the general theory. It comprises all $(n+1) \times (n+1)$ matrices of trace zero, and splits as a direct sum in the following way:

$$\mathfrak{sl}_{n+1} = \mathfrak{n}_+ \oplus \mathfrak{h} \oplus \mathfrak{n}_-,$$

where \mathfrak{h} is the set of diagonal matrices of trace zero, and \mathfrak{n}_+ and \mathfrak{n}_- are, respectively, the sets of upper and lower triangular matrices with 0s on the diagonal. Two diagonal matrices X and Y commute with one another, so their Lie bracket $[X, Y] = XY - YX$ is 0. In other words, if X and Y belong to \mathfrak{h} , then $[X, Y] = 0$. A Lie algebra in which $[X, Y] = 0$ for any two elements X and Y is called *Abelian*.

Each simple Lie algebra \mathfrak{g} has a similar decomposition where the subspace \mathfrak{h} is a maximal Abelian subalgebra called a *Cartan subalgebra*. (For Lie algebras that are not simple, the definition of Cartan subalgebras is more complicated.) Cartan subalgebras are important because their action on the rest of the Lie algebra can be simultaneously diagonalized. What this means is that a complement to \mathfrak{h} can be split up into one-dimensional components \mathfrak{g}_α , known as *root spaces*, that are invariant under the action of \mathfrak{h} . To put this another way, if X belongs to \mathfrak{h} , and Y belongs to a root space, then $[X, Y]$

1. Indeed, Lie groups and Lie algebras are an excellent tool for describing the algebraic aspects of ordinary and partial differential equations; the evolution of such equations through time can be modeled using a Lie group, and the differential operators used to describe an equation can be modeled on the associated Lie algebra. However, we will not discuss this important connection between Lie theory and differential equations here.

is a scalar multiple of Y . (The diagonalization requires THE FUNDAMENTAL THEOREM OF ALGEBRA [V.15], which is why we need to work with complex Lie algebras.)

For \mathfrak{sl}_{n+1} this works as follows. Each root space \mathfrak{g}_{ij} is the one-dimensional space of matrices whose entries are 0 except for a single entry in the i th row and j th column. If $X \in \mathfrak{h}$ (that is, if X is a diagonal matrix of trace zero) and $Y \in \mathfrak{g}_{ij}$, then it is not hard to check that $[X, Y]$ also lies in \mathfrak{g}_{ij} . In fact,

$$[X, Y] = (X_{ii} - X_{jj})Y.$$

If we identify the diagonal matrix X with the vector whose n coordinates appear down its diagonal, and if we write e_i for the vector that is 1 in the i th position and 0 elsewhere, then $X_{ii} - X_{jj}$ can be rewritten as $\langle e_i - e_j, X \rangle$. We refer to the vectors $e_i - e_j$ as *root vectors*.

In general, a **complex semisimple** Lie algebra \mathfrak{g} can be completely described by its root vectors α and corresponding root spaces \mathfrak{g}_α . The *rank* of \mathfrak{g} equals the dimension of the Cartan subalgebra \mathfrak{h} , and also equals the dimension of the vector space spanned by the root vectors. For example, \mathfrak{sl}_{n+1} has rank n , and its root vectors are the vectors $e_i - e_j$, as we have just seen. Sets of root vectors are far from arbitrary: they must obey some simple but quite restrictive geometric properties. For instance, if a root vector α is reflected in the hyperplane perpendicular to another root vector β , the result must be a third root vector $s_\beta(\alpha)$, where s_β is the reflection concerned. (To make the notion of “perpendicular” precise, one needs to define a special inner product on the Cartan subalgebra, known as the *Killing form*, but we shall not discuss this here.) The group generated by these reflections is called the *Weyl group* of the Lie algebra.

The root vectors form what is called a *root system*, and the geometric properties mentioned above allow one to classify all root systems, and hence all complex semisimple Lie algebras. This classification is given by some very simple diagrams called *Dynkin diagrams*, which are shown in figure 1.

The nodes of the diagram correspond to so-called *simple roots*. Every root is a linear combination of simple roots with coefficients that are either all nonnegative or all nonpositive. The nature of the bond (or lack thereof) between two nodes determines the inner product of the corresponding simple roots. If there is no bond, then the inner product is 0; if there is a single bond, then the root vectors have the same length and the angle between them is 120° . In diagrams that have

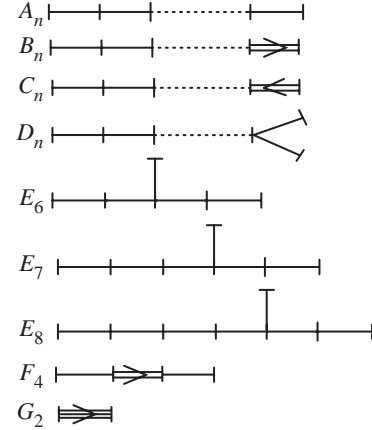


Figure 1 Dynkin diagrams.

only single bonds, the root vectors span a set of lines in \mathbb{R}^n in which the angle between any two lines is either 90° or 60° . In the diagrams B_n , C_n , F_4 , and G_2 there are arrows between certain pairs of nodes. The direction of an arrow is from a long root to a short root: the ratio of the root lengths is $\sqrt{2}$ in the first three cases and $\sqrt{3}$ in the case of G_2 . In these cases there are exactly two root lengths, but in the single-bond cases all roots have the same length.

The A_n diagram is the one for \mathfrak{sl}_{n+1} . The simple roots are $e_i - e_{i+1}$ for $1 \leq i \leq n$, going from left to right on the diagram. Notice that the inner product of two simple roots is 0 unless they are adjacent on the diagram, in which case it is -1 . Each root $e_i - e_j$ is a sum of simple roots with coefficients all 1 or all -1 on a connected segment of the diagram.

The four infinite families A_n , B_n , C_n , and D_n correspond to the *classical Lie algebras*, of which $\mathfrak{sl}_{n+1}(\mathbb{R})$, $\mathfrak{so}(2n+1)$, $\mathfrak{sp}(2n)$, and $\mathfrak{so}(2n)$ are real forms. These are the algebras associated with the *classical Lie groups* $SL_n(\mathbb{R})$, $SO(2n+1)$, $Sp(2n)$, and $SO(2n)$, respectively.

As mentioned earlier, a simple Lie algebra \mathfrak{g} of rank n decomposes as the direct sum of a Cartan subalgebra of dimension n plus a set of one-dimensional root spaces, one for each root. It follows that

$$\dim \mathfrak{g} = \text{the rank of } \mathfrak{g} + \text{the number of roots}.$$

Here are the dimensions of the simple Lie algebras:

$$\dim A_n = n + n(n+1) = n(n+2),$$

$$\dim B_n = n + 2n^2 = n(2n+1),$$

$$\dim C_n = n + 2n^2 = n(2n+1),$$

$$\dim D_n = n + 2n(n-1) = n(2n-1),$$

PUP: complex is being used in a mathematical sense here and all the other times when the proofreader queried the combination of 'complex' and 'simple'. They are all OK.

T&T note: Tim checking with author whether 'n' should be 'n+1' here.

$$\dim G_2 = 2 + 12 = 14,$$

$$\dim F_4 = 4 + 48 = 52,$$

$$\dim E_6 = 6 + 72 = 78,$$

$$\dim E_7 = 7 + 126 = 133,$$

$$\dim E_8 = 8 + 240 = 248.$$

Each node of the diagram corresponds to a simple root, and hence to a reflection across the hyperplane perpendicular to that root. This set of reflections generates the Weyl group W in a particularly elegant way. If s_i denotes the reflection corresponding to node i , then W is generated by elements s_i of order 2, subject only to the relations

$$(s_i s_j)^{m_{ij}} = 1,$$

where m_{ij} is the order of $s_i s_j$ (see [IV.11 §2] for a discussion of generators and relations). These orders are determined by the diagram according to the following rules:

- (i) $s_i s_j$ has order 2 if there is no bond;
- (ii) $s_i s_j$ has order 3 if there is a single bond;
- (iii) $s_i s_j$ has order 4 if there is a double bond; and
- (iv) $s_i s_j$ has order 6 if there is a triple bond.

For example, the Weyl group of type A_n is isomorphic to the SYMMETRIC GROUP [III.70] S_{n+1} , and one can take s_1, \dots, s_n to be the transpositions $(1\ 2), (2\ 3), \dots, (n\ n+1)$. Notice that the Dynkin diagrams for the B_n and C_n root systems yield the same Weyl group.

In principle, this classification of root systems leads to a classification of all semisimple finite-dimensional Lie algebras and Lie groups. However, there are many fundamental questions about simple Lie algebras and Lie groups that remain only partly understood. For instance, one particularly important aim of Lie theory is to understand the linear representations of a given Lie group or Lie algebra; roughly speaking, a linear representation is a way of interpreting an abstract Lie group or Lie algebra as a linear Lie group or Lie algebra by assigning a matrix to each of its elements. While the representations of all the simple Lie algebras and Lie groups have been classified and described explicitly, these descriptions are not always easy to work with, and answering basic questions (such as how a given representation decomposes into simpler representations) often requires some sophisticated tools from algebraic combinatorics.

The theory of root systems outlined above can also be extended to an important class of infinite-dimensional Lie algebras, namely the Kac-Moody algebras. Such

algebras arise in several areas of physics (such as are described in VERTEX OPERATOR ALGEBRAS [IV.13]) and algebraic combinatorics.

III.51 Linear and Nonlinear Waves and Solitons

Richard S. Palais

1 John Scott Russell and the Great Wave of Translation

To the world at large, John Scott Russell is known as the naval architect who designed *The Great Eastern*, a steamship larger than any built before. But long after *The Great Eastern* has been forgotten, Russell will be remembered by mathematicians as the man who, despite limited mathematical training and background, was the first person to recognize the highly important mathematical concept known as a *soliton*, which he referred to as “the great wave of translation.” Here is his oft-quoted passage in which he describes how he first became acquainted with it:

I was observing the motion of a boat which was rapidly drawn along a narrow channel by a pair of horses, when the boat suddenly stopped—not so the mass of water in the channel which it had put in motion; it accumulated round the prow of the vessel in a state of violent agitation, then suddenly leaving it behind, rolled forward with great velocity, assuming the form of a large solitary elevation, a rounded, smooth and well-defined heap of water, which continued its course along the channel apparently without change of form or diminution of speed. I followed it on horseback, and overtook it still rolling on at a rate of some eight or nine miles an hour, preserving its original figure some thirty feet long and a foot to a foot and a half in height. Its height gradually diminished, and after a chase of one or two miles I lost it in the windings of the channel. Such, in the month of August 1834, was my first chance interview with that singular and beautiful phenomenon which I have called the Wave of Translation.

Russell (1844)

You may feel that there is nothing unusual about what Russell describes here, and indeed many before and since have watched this same scenario play out without noticing anything out of the ordinary. But Russell was very familiar with wave phenomena and had a scientist’s keenly observant eye. What struck him was the remarkable *stability* of the bow wave as it traveled over a long distance. He knew that if one tried to create a traveling water wave on, say, a calm lake, it would

soon disperse into a train of smaller wavelets—it would *not* just go marching along as a single “heap” over a long distance. There was clearly something very special about water waves traveling in a narrow and shallow channel.

Russell became fascinated—even a little obsessed—with his discovery. He built a wave tank behind his home and proceeded to do extensive experiments, recording the results as data and sketches in his notebooks. He found, for example, that the speed of a soliton depended on its height, and he was even able to discover the correct formula for the speed as a function of height. More surprising still, in Russell’s notebooks one finds remarkable sketches of a two-soliton interaction—something that would evoke amazement when it was rediscovered as a rigorous solution to the KdV equation (see section 3 below) more than a hundred years later.

However, as we shall see, solitons are very much a nonlinear phenomenon, and when some of the best mathematicians of Russell’s day, notably Stokes and Airy, tried to understand Russell’s observations using the linearized theory of water waves that was then available, they failed to find any trace of soliton-like behavior and expressed doubts that what Russell had seen was real.

It was only after Russell’s death, with the more sophisticated nonlinear mathematical treatment by Boussinesq in 1871 and by Korteweg and de Vries in 1895, that Russell’s careful observations and experiments were at last seen to be in complete agreement with mathematical theory. And it took another seventy years before the full importance of the great wave of translation was recognized, after which it became an object of intensive study for the rest of the twentieth century.

2 The Korteweg–de Vries Equation

Korteweg and de Vries were the first to derive the appropriate differential equation to describe the motion of a wave in a shallow channel. We can write their equation, usually called the *KdV equation*, in a succinct form as follows:

$$u_t + uu_x + \delta^2 u_{xxx} = 0.$$

Here, u is a function of two variables, x and t , which represent space and time, respectively. “Space” is one dimensional, so x is a real number, and $u(x, t)$ represents the height of the wave at x at time t . The nota-

tion u_t is shorthand for $\partial u / \partial t$; similarly, u_x stands for $\partial u / \partial x$ and u_{xxx} stands for $\partial^3 u / \partial x^3$.

This is an example of an *evolution equation*: if, for each t , we write $u(t)$ for the function from \mathbb{R} to \mathbb{R} that takes x to $u(x, t)$, then it describes how the function $u(t)$ “evolves” over time. The *Cauchy problem* for an evolution equation is the problem of determining this evolution from knowledge of its initial value $u(0)$.

2.1 Some Model Equations

To put the KdV equation into perspective, it is useful to think briefly about three other evolution equations. The first is the classic WAVE EQUATION [I.3 §5.4]

$$u_{tt} - c^2 u_{xx} = 0.$$

To solve the Cauchy problem for this equation, we factor the wave operator $(\partial^2 / \partial t^2) - c^2 (\partial^2 / \partial x^2)$ as a product $((\partial / \partial t) - c(\partial / \partial x))((\partial / \partial t) + c(\partial / \partial x))$. Then we transform to so-called characteristic coordinates $\xi = x - ct$, $\eta = x + ct$. The equation becomes $\partial^2 u / \partial \xi \partial \eta = 0$, which clearly has the general solution $u(\xi, \eta) = F(\xi) + G(\eta)$. Transforming back to “laboratory coordinates” x, t , the general solution is $u(x, t) = F(x - ct) + G(x + ct)$. If the initial shape of the wave is $u(x, 0) = u_0(x)$ and its initial velocity is $u_t(x, 0) = v(x, 0) = v_0(x)$, then an easy algebraic computation gives the following very explicit formula:

$$u(x, t) = \frac{1}{2} [u_0(x - ct) + u_0(x + ct)] + \frac{1}{2c} \int_{x-ct}^{x+ct} v_0(\xi) d\xi,$$

which is known as “d’Alembert’s solution” of the Cauchy problem for the wave equation.

Note the geometric interpretation in the important “plucked string” case, $v_0 = 0$; the initial profile u_0 breaks up into the sum of two “traveling waves,” both with the same profile $\frac{1}{2}u_0$, one traveling to the right, and the other to the left, both with speed c . It is an easy exercise to derive d’Alembert’s solution using the following hint: since $u_0(x) = F(x) + G(x)$, $u'_0(x) = F'(x) + G'(x)$, while $v_0(x) = u_t(x, 0) = -cF'(x) + cG'(x)$.

The next equation to think about is

$$u_t = -u_{xxx}, \quad (1)$$

which we can obtain from the KdV equation if we drop the nonlinear term uu_x . This equation is not just linear but also translation invariant (meaning that if $u(x, t)$ is a solution, then so is $u(x - x_0, t - t_0)$ for any constants x_0 and t_0). Such equations can be solved using

THE FOURIER TRANSFORM [III.27]. Let us try to find a “plane-wave” solution of the form $u(x, t) = e^{i(kx - \omega t)}$. If we substitute this into (1), then we obtain the equation

$$-i\omega e^{i(kx - \omega t)} = ik^3 e^{i(kx - \omega t)},$$

and therefore the simple algebraic equation $\omega + k^3 = 0$. This is called the *dispersion relation* of (1): with the help of the Fourier transform it is not hard to show that every solution is a superposition of solutions of the form $e^{i(kx - \omega t)}$, and the dispersion relation tells us how the “wave number” k is related to the “angular frequency” ω in each of these elementary solutions.

The function $e^{i(kx - \omega t)}$ represents a wave that travels at a speed of ω/k , which we have just shown to be equal to $-k^2$. Therefore, the different plane-wave components of the solution travel at different speeds: the higher the angular frequency, the greater the speed. For this reason, the equation (1) is called *dispersive*.

What happens if instead we omit the u_{xxx} term from the KdV equation? Then we obtain the *inviscid Burgers equation*

$$u_t + uu_x = 0. \quad (2)$$

The term uu_x can be rewritten as $(\partial/\partial x)(\frac{1}{2}u^2)$. Let us consider the integral $\int_{-\infty}^{\infty} u(x, t) dx$, which is a function of t . The derivative of this function is $\int_{-\infty}^{\infty} u_t dx$, which equation (2) tells us is equal to

$$-\int_{-\infty}^{\infty} \frac{\partial}{\partial x} \left(\frac{1}{2} u^2 \right) dx,$$

which equals $[-\frac{1}{2}u(x, t)^2]_{-\infty}^{\infty}$. Therefore, if $\frac{1}{2}u(x, t)^2$ vanishes at infinity, then the original expression $\int_{-\infty}^{\infty} u(x, t) dx$ is a “constant of the motion.” We say that the inviscid Burgers equation is a *conservation law*. (The argument we have just used can be used for any equation of the form $u_t = (F(u))_x$, where F is a smooth function of u and its partial derivatives with respect to x . This is known as the *general conservation law*. For example, taking $F(u) = -(\frac{1}{2}u^2 + \delta^2 u_{xx})$ gives rise to the KdV equation.)

The inviscid Burgers equation (and other conservation laws where F is a function just of u) can be solved using the *method of characteristics*. The idea of this method is to look for smooth curves $(x(s), t(s))$ in the xt -plane along which the solution to the Cauchy problem is constant. Suppose that s_0 is such that $t(s_0) = 0$, and write x_0 for $x(s_0)$. Then the constant value that the solution $u(x, t)$ will have to take along this curve is $u(x_0, 0)$, which we also write as $u_0(x_0)$. The derivative of u along this so-called *characteristic curve* is $(d/ds)u(x(s), t(s)) = u_x x' + u_t t'$, so if we want the solution to be constant along the curve, then we need

this to be 0. Therefore, using the fact that $u_t = -uu_x$, we find that

$$\frac{dx}{dt} = \frac{x'(s)}{t'(s)} = -\frac{u_t}{u_x} = u(x(s), t(s)) = u_0(x_0),$$

so the characteristic curve is a straight line of slope $u_0(x_0)$. In other words, u has the constant value $u_0(x_0)$ along the line $x = x_0 + u_0(x_0)t$.

Note the following geometric interpretation of this last result: to find the wave profile at time t (i.e., the graph of the map $x \mapsto u(x, t)$), we translate each point $(x, u_0(x))$ of the initial profile to the right by the amount $u_0(x)t$. Suppose we look at a portion of the initial profile where u_0 is decreasing. Then the earlier, and higher, parts of the initial wave are translated at a greater speed (since $u_0(x)$ is larger), so that the negative slope of the wave becomes more negative. Indeed, after a finite time the earlier part of the wave “catches up” with the later part, which means that we no longer have a graph of a function. The first time at which this sort of problem happens is called the “breaking time,” since one can visualize it as the breaking of a wave. This process is usually referred to as *shock formation*, or *steepening and breaking of the wave profile*: once again, the phenomenon occurs for many other conservation laws.

2.2 Split-Stepping

Now let us return to the KdV equation itself, in the form $u_t = -uu_x - u_{xxx}$. Why is it that this equation gives rise to the remarkable stability of the solutions that was observed experimentally by Russell? Intuitively, the reason is that there is a balance between the dispersing effect of the u_{xxx} term and the shock-forming effect of the uu_x term.

There turns out to be a very general technique for analyzing balances of this kind. In the pure-mathematics community it is usually called the *Trotter product formula*, while in the applied-mathematics and numerical-analysis communities it is called *split-stepping*. The rough idea is simple: as t increases to $t + \Delta t$, you first change u to $u - u_{xxx}\Delta t$, as would be required by the equation $u_t = -u_{xxx}$, and then you take a further step to $u - u_{xxx}\Delta t - uu_x\Delta t$, the small change required by the equation $u_t = -uu_x$. To work out the function $u(t, x)$, you start at the initial function u_0 and take a succession of alternating small steps of this form. You then take the limit as the step size tends to zero.

Split-stepping suggests a way to understand the mechanism by which dispersion from u_{xxx} balances

shock formation from uu_x in KdV. If we imagine the evolution of the wave profile as made up of a succession of pairs of small steps in this way, then when u , u_x , and u_{xxx} are not too large, the steepening mechanism will dominate. But as the time t approaches the breaking time T_B , u remains bounded (since it is made out of horizontally translated parts of u_0). It is not hard to prove that the maximum slope (that is, the maximum value of u_x) blows up like the function $(T_B - t)^{-1}$, while at the same place, u_{xxx} blows up like the function $(T_B - t)^{-5}$. Thus, near the breaking time, and breaking point, the u_{xxx} term will dwarf the nonlinearity and will disperse the incipient shock. Thus, the stability is caused by a kind of negative feedback. Computer simulations show just such a scenario playing out.

3 Solitons and Their Interactions

We have just seen that the KdV equation expresses a balance between dispersion from its third-derivative term and the shock-forming tendency of its nonlinear term, and in fact many models of one-dimensional physical systems that exhibit mild dispersion and weak nonlinearity lead to KdV as the controlling equation at some level of approximation.

In their 1894 paper, Korteweg and de Vries introduced the KdV equation and gave a convincing mathematical argument that this was the equation that governed wave motion in a shallow canal. They also showed by explicit computation that it admitted traveling-wave solutions that had exactly the properties that had been described by Russell, including the relation of height to speed that Russell had determined experimentally with the help of his wave tank.

But it was only much later that further remarkable properties of the KdV equation became evident. In 1954, Fermi, Pasta, and Ulam (FPU) used one of the very first digital computers to perform numerical experiments on an elastic string with a nonlinear restoring force, and their results contradicted the then current expectations of how energy should distribute itself among the normal modes of such a system. A decade later, Zabusky and Kruskal reexamined the FPU results in a famous paper in which they showed that the FPU string was well approximated by the KdV equation. They then did their own computer experiments, solving the Cauchy problem for KdV with initial conditions corresponding to those used in the FPU experiments. In the results of these simulations they observed the first example of a “soliton,” a term that they coined

to describe a remarkable particle-like behavior (elastic scattering) exhibited by certain KdV solutions. Zabusky and Kruskal showed how the coherence of solitons explained the anomalous results observed by Fermi, Pasta, and Ulam. But in solving that mystery they had uncovered a larger one: the behavior of KdV solitons was unlike anything seen before in applied mathematics, and the search for an explanation of their remarkable behavior led to a series of discoveries that changed the course of applied mathematics for the next thirty years. We shall now fill in some of the mathematical details behind the above sketch, beginning with a discussion of explicit solutions to the KdV equation.

To find the traveling-wave solutions of KdV is straightforward. First, we substitute a traveling wave $u(x, t) = f(x - ct)$ into KdV, obtaining the ordinary differential equation $-cf' + 6ff' + f''' = 0$. If we add as a boundary condition that f should vanish at infinity, then a routine computation leads to the following two-parameter family of traveling-wave solutions:

$$u(x, t) = 2a^2 \operatorname{sech}^2(a(x - 4a^2t + d)).$$

These are the solitary waves seen by Russell, and they are now usually referred to as the 1-soliton solutions of KdV. Note that their amplitude, $2a^2$, is just half their speed, $4a^2$, while their “width” is proportional to a^{-1} . Thus, taller solitary waves are thinner and move faster.

Next, following Toda, we will “derive”¹ the 2-soliton solutions of KdV. Rewrite the 1-soliton solution as $u(x, t) = 2(\partial^2/\partial x^2) \log \cosh(a(x - 4a^2t + \delta))$, or $u(x, t) = 2(\partial^2/\partial x^2) \log K(x, t)$, where $K(x, t) = (1 + e^{2a(x - 4a^2t + \delta)})$. We now try to generalize, looking for solutions of the form $u(x, t) = 2(\partial^2/\partial x^2) \log K(x, t)$, with $K(x, t) = 1 + A_1 e^{2\eta_1} + A_2 e^{2\eta_2} + A_3 e^{2(\eta_1 + \eta_2)}$, where $\eta_i = a_i(x - 4a_i^2t + d_i)$, and we shall choose the A_i and d_i by substituting into KdV and seeing what works. One can check that KdV is satisfied for $u(x, t)$ of this form and arbitrary $A_1, A_2, a_1, a_2, d_1, d_2$, provided that we define $A_3 = ((a_2 - a_1)/(a_1 + a_2))^2 A_1 A_2$, and solutions of KdV arising in this way are called the KdV 2-soliton solutions.

It can now be shown that for these choices of a_1 and a_2 ,

$$u(x, t) = 12 \frac{3 + 4 \cosh(2x - 8t) + \cosh(4x - 64t)}{[\cosh(3x - 36t) + 3 \cosh(x - 28t)]^2}.$$

In particular, $u(x, 0) = 6 \operatorname{sech}^2(x)$, $u(x, t)$ is asymptotically equal to $2 \operatorname{sech}^2(x - 4t - \phi) + 8 \operatorname{sech}^2(x - 16t + \frac{1}{2}\phi)$

1. This is a complete swindle! Only knowledge of the form of the solutions allows us to make the clever choice of K .

when t is large and negative, and $u(x, t)$ is asymptotically equal to $2 \operatorname{sech}^2(x - 4t + \phi) + 8 \operatorname{sech}^2(x - 16t - \frac{1}{2}\phi)$ when t is large and positive, where $\phi = \frac{1}{3} \log(3)$.

Note what this says. If we follow the evolution from $-T$ to T (where T is large and positive), we first see the superposition of two 1-solitons: a larger and thinner one to the left of, and catching up with, a shorter, fatter, and slower-moving one to the right. Around $t = 0$ they merge into a single lump (with the shape $6 \operatorname{sech}^2(x)$), and then they separate again, with their original shapes restored—but now the taller and thinner one is to the right. It is almost as if they had passed right through each other. The only effect of their interaction is the pair of phase shifts: the slower one is retarded slightly from where it would have been, and the faster one is slightly ahead of where it would have been. Except for these phase shifts, the final result is what we might expect from a linear interaction. It is only if we look closely at the interaction as the two solitons meet that we can detect its highly nonlinear nature. (Note, for example, that at time $t = 0$, the maximum amplitude, 6, of the combined wave is actually less than the maximum amplitude, 8, of the taller wave when they are separated.) But of course the really striking fact is the resilience of the two individual solitons: their ability to put themselves back together after the collision. Not only is no energy radiated away, but their actual shapes are preserved. (Remarkably, Russell (1844, p. 384) gives a sketch of a 2-soliton interaction experiment that he had carried out in his wave tank!)

Now back to the computer experiment of Zabusky and Kruskal. For numerical reasons, they chose to deal with the case of periodic boundary conditions: in effect, studying the KdV equation $u_t + uu_x + \delta^2 u_{xxx} = 0$ (which they label (1)) on the circle instead of on the line. For their published report, they chose $\delta = 0.022$ and used the initial condition $u(x, 0) = \cos(\pi x)$. With the above background in mind, it is interesting to read the following extract from their 1965 report, which contains the first use of the term “soliton”:

(I) Initially the first two terms of Eq. (1) dominate and the classical overtaking phenomenon occurs; that is u steepens in regions where it has negative slope. (II) Second, after u has steepened sufficiently, the third term becomes important and serves to prevent the formation of a discontinuity. Instead, oscillations of small wavelength (of order δ) develop on the left of the front. The amplitudes of the oscillations grow, and finally *each* oscillation achieves an almost steady amplitude (that increases linearly from left to right) and has the shape of an individual solitary-wave of (1). (III) Finally,

each “solitary wave pulse” or *soliton* begins to move uniformly at a rate (relative to the background value of u from which the pulse rises) which is linearly proportional to its amplitude. Thus, the solitons spread apart. Because of the periodicity, two or more solitons eventually overlap spatially and interact nonlinearly. Shortly after the interaction they reappear virtually unaffected in size or shape. In other words, solitons “pass through” one another without losing their identity. *Here we have a nonlinear physical process in which interacting localized pulses do not scatter irreversibly.*

Zabusky and Kruskal (1965)

Further Reading

- Lax, P. D. 1996. *Outline of a Theory of the KdV Equation in Recent Mathematical Methods in Nonlinear Wave Propagation*. Lecture Notes in Mathematics, volume 1640, pp. 70–102. New York: Springer.
- Palais, R. S. 1997. The symmetries of solitons. *Bulletin of the American Mathematical Society* 34:339–403.
- Russell, J. S. 1844. Report on waves. In *Report of the 14th Meeting of the British Association for the Advancement of Science*, pp. 311–90. London: John Murray.
- Toda, M. 1989. *Nonlinear Waves and Solitons*. Dordrecht: Kluwer.
- Zabusky, N. J., and M. D. Kruskal. 1965. Interaction of solitons in a collisionless plasma and the recurrence of initial states. *Physics Review Letters* 15:240–43.

III.52 Linear Operators and Their Properties

1 Some Examples of Linear Operators

A LINEAR MAP [I.3 §4.2] between VECTOR SPACES [I.3 §2.3] V and W is a function $T : V \rightarrow W$ that satisfies the condition $T(\lambda_1 v_1 + \lambda_2 v_2) = \lambda_1 T v_1 + \lambda_2 T v_2$. Two phrases that are used almost interchangeably with “linear map” are “linear transformation” and “linear operator.” The former is often used when one wishes to draw attention to the effect of a linear map on some other object; for example, one might well choose to use the word “transformation” to describe geometrical operations such as reflections or rotations. As for “operator,” it tends to be the word of choice when the linear map is between infinite-dimensional spaces, especially when it is just one of an ensemble of linear maps that form an algebra. It is these maps that we shall discuss here.

Let us begin with some examples of linear operators.

(i) If X is a BANACH SPACE [III.64] whose elements are infinite sequences, then we can define a “shift” S from X to X , which takes the sequence (a_1, a_2, a_3, \dots) to the sequence $(0, a_1, a_2, a_3, \dots)$. (In other words, it puts a 0 at the beginning and shifts the other values of the sequence one place to the right.) The map S is linear, and if the norm on X is not too pathological, then S will be a continuous function from X to X .

(ii) If X is a SPACE OF FUNCTIONS [III.29] defined on the closed interval $[0, 1]$ and w is some fixed function, then the map M that takes the function f to the product fw (which is shorthand for the function $x \mapsto f(x)w(x)$) is linear, and, provided w is small enough in some appropriate sense, M is a continuous linear map from X to X . Such maps are called *multipliers*. (Note that the property of “being a multiplier” depends not just on the space X and the map M but also on the way we choose to represent X as a space of functions, so it is not an intrinsic property of the map itself.)

(iii) Another important way of defining linear operators on function spaces is to use a *kernel*. This is a function K of two variables, which can be used to define a linear map in a way that is similar to the way a matrix can be used to define a map between finite-dimensional vector spaces. The following formula uses K to define a linear map T :

$$Tf(x) = \int K(x, y)f(y) dy. \quad (1)$$

Note the formal similarity between this and the formula

$$(Av)_i = \sum_j A_{ij}v_j,$$

which defines the product of a matrix with a column vector. Once again, K will have to satisfy appropriate conditions in order for (1) to define a continuous linear map.

A good example of a linear operator defined by a kernel is THE FOURIER TRANSFORM [III.27] \mathcal{F} , which takes a function in $L^2(\mathbb{R})$ to another such function. It is defined by the formula

$$(\mathcal{F}f)(\alpha) = \int_{-\infty}^{\infty} f(x)e^{-i\alpha x} dx.$$

The kernel in this case is the function $K(\alpha, x) = e^{-i\alpha x}$.

(iv) If f is a differentiable function defined on \mathbb{R} , say, and we write Df for its derivative, then we can think of D as a linear map, since $D(\lambda f + \mu g) = \lambda Df + \mu Dg$. In order to regard D as an operator, we need to require f to belong to a suitable function space. The best way of doing this varies from context to context: choosing a good function space can be very important and

can raise subtle questions. One way is not to insist that D is defined for every function in the space, but just on a dense set of functions, and not to require that D is continuous. Similarly, many partial differential operators, such as the GRADIENT [I.3 §5.3] and the LAPLACIAN [I.3 §5.4], are linear operators when viewed appropriately.

2 Algebras of Operators

Although individual operators can be important, linear operators would not be as interesting as they are if it were not for the fact that they can be formed into *families*. If X is a Banach space, then the set $B(X)$ of all continuous linear operators from X to itself forms a structure known as a *Banach algebra*. Roughly speaking, this means that it is a Banach space (the norm of an operator T is defined to be the supremum of $\|Tx\|$ over all x such that $\|x\| \leq 1$) in which the elements can be multiplied as well as added. The product of T_1 and T_2 is defined to be the composition $T_1 T_2$, and it is easily seen to satisfy the inequality $\|T_1 T_2\| \leq \|T_1\| \|T_2\|$. This algebra is particularly important when X is a HILBERT SPACE [III.37] H : subalgebras of $B(H)$ have a very rich structure, which is discussed in OPERATOR ALGEBRAS [IV.19].

3 Properties of Operators Defined on a Hilbert Space

Unlike a general Banach space, a Hilbert space H has an inner product. It is therefore natural to ask that a continuous linear operator from H to H should relate to the inner product somehow. This basic idea leads to several different definitions, each of which picks out an important class of operators.

3.1 Unitary and Orthogonal Maps

Perhaps the most obvious condition one might require of an operator T is that it should *preserve* the inner product, in the sense that $\langle Tx, Ty \rangle$ should equal $\langle x, y \rangle$ for any two vectors x and y . In particular, this implies that $\|Tx\| = \|x\|$ for every x , and therefore that T is an *isometry* (that is, a map that preserves distances). If in addition, T is invertible, which it will be if its image is the whole of H , then T is a *unitary* map. The unitary maps form a group. If H is n dimensional, then this group is an important LIE GROUP [III.50 §1] called $U(n)$. If H is a real Hilbert space (as opposed to a complex one), then the word “orthogonal” is used instead

of “unitary” and the corresponding Lie group is called $O(n)$. When $n = 3$, orthogonal maps are rotations and reflections, so $O(n)$ is the generalization of the group of rotations and reflections to n dimensions.

3.2 Hermitian and Self-Adjoint Maps

Given any operator T from H to H , there is an operator T^* from H to H with the property that $\langle Tx, y \rangle = \langle x, T^*y \rangle$ for every x and y . This operator is unique, and it is called the *adjoint* of T . A second property that T can have is that of equaling its own adjoint, which is the case if and only if $\langle Tx, y \rangle = \langle x, Ty \rangle$ for every x and y . Such operators are called *Hermitian* or, when the scalars are real, *self-adjoint*. A simple source of examples of Hermitian maps is multipliers on the space $L^2[0, 1]$, where the function one multiplies by is bounded and real-valued. As we shall see in a moment, there is a sense in which these are the only examples.

3.3 Properties of Matrices

If H is a finite-dimensional space with an orthonormal basis, then we can form the matrix A of T with respect to that basis. The various properties of T discussed above then turn out to be equivalent to properties of the matrix A . The *transpose* of A is the matrix A^T defined by $(A^T)_{ij} = A_{ji}$, and the *conjugate transpose* is the matrix A^* defined by $(A^T)_{ij} = \overline{A_{ji}}$. An $n \times n$ matrix is *unitary* if AA^* is the identity, *orthogonal* if it is real and AA^T is the identity, *Hermitian* if $A = A^*$, and *self-adjoint* if $A = A^T$ (in which case we say that A is *symmetric*). The operator T has one of these four properties if and only if its matrix A has the corresponding property.

3.4 The Spectral Theorem

Notice that the adjoint of a unitary operator is the *inverse* of that operator. In particular, both unitary and Hermitian operators commute with their adjoints. An operator with this property is called *normal*. Normal operators are important because of the famous spectral theorem. If T is a normal operator on a finite-dimensional space H , then the spectral theorem asserts that H has an ORTHONORMAL BASIS [III.37] of eigenvectors of T . In other words, there is a basis of H consisting of orthogonal unit vectors, with the property that the matrix of T with respect to this basis is diagonal. This is an extremely useful theorem in linear algebra. In general, if T is a normal operator on a Hilbert space H , then the spectral theorem tells us that there is something

like a “basis” for H , with respect to which T is a multiplier. To put this slightly differently, there is an isometric isomorphism ϕ from H to a Hilbert space H' of functions that are square-integrable with respect to some MEASURE [III.57], and the map $\phi T \phi^{-1}$ is a multiplier on H' .

3.5 Projections

Another important class of maps on a Hilbert space is the set of *orthogonal projections*. In general, an element T of an algebra is an *idempotent* if it has the property that $T^2 = T$. If the algebra is an algebra of operators on a space X , then T is called a *projection*. To see why this name is appropriate, note that every x is mapped to the subspace TX of X , and all points in that subspace are left fixed by T (since $T(Tx) = T^2x = Tx$). A projection is *orthogonal* if Tx is always orthogonal to $x - Tx$. This tells us that T is a projection on to some subspace Y of H , and that it takes each vector to the nearest point in Y , so that the vector $x - Tx$ is orthogonal to the whole of the subspace Y .

III.53 Local and Global in Number Theory

Fernando Q. Gouvêa

Analogy is a powerful tool. When one can see parallels between two different theories, this often allows one to transport insights from one to the other. The idea of studying something “locally” comes from the theory of functions. Imported into number theory by way of an analogy between functions and numbers, it leads us to a whole new kind of number, the p -adic numbers, and to the *local-global principle*, which has become one of the guiding ideas of modern number theory.

1 Studying Functions Locally

Suppose that we have a polynomial such as

$$f(x) = -18 + 21x - 26x^2 + 22x^3 - 8x^4 + x^5.$$

From the very way the polynomial is written down, we can see certain things about it. For example, we can see at once that if we plug in $x = 0$, we get $f(0) = -18$. Other things are less apparent. For example, to decide what $f(2)$ or $f(3)$ are, we would have to do some arithmetic. But if we were to rewrite the polynomial as

$$\begin{aligned} f(x) = & 5(x-2) - 6(x-2)^2 - 2(x-2)^3 \\ & + 2(x-2)^4 + (x-2)^5, \end{aligned}$$

we could see at once that $f(2) = 0$. (Of course, one needs to check that those two expressions really are equal!) Similarly, we can check that

$$f(x) = 10(x-3)^2 + 16(x-3)^3 + 7(x-3)^4 + (x-3)^5$$

and see at once that $f(3)$ is also zero, and in fact that the polynomial has a double root at $x = 3$.

One way to think about this is to describe the first expression as “local at $x = 0$,” because it privileges the value 0 over all others. Then the other two expressions are local at 2 and local at 3, respectively. On the other hand, a formula like

$$f(x) = (x-2)(x-3)^2(x^2+1)$$

(which is also correct) is clearly more “global.” It tells us where all the roots are: at 2, 3, and $\pm\sqrt{-1}$, with the 3 being a double root.

The same ideas extend to functions that are not polynomials, as long as we allow the expressions to be infinite. So, for example, let us take

$$g(x) = \frac{x^2 - 5x + 2}{x^3 - 2x^2 + 2x - 4}.$$

Locally at 0, we can write this as

$$g(x) = -\frac{1}{2} + x + \frac{1}{2}x^2 - \frac{3}{8}x^3 - \frac{3}{16}x^4 + \frac{7}{32}x^5 + \dots$$

Or we can write it locally at 2:

$$\begin{aligned} g(x) &= -\frac{2}{3}(x-2)^{-1} + \frac{5}{18} + \frac{5}{54}(x-2) \\ &\quad - \frac{35}{324}(x-2)^2 + \frac{55}{972}(x-2)^3 \\ &\quad - \frac{115}{5832}(x-2)^4 + \frac{65}{17496}(x-2)^5 + \dots \end{aligned}$$

Notice that this time we had to use a *negative* power of $(x-2)$, because plugging in $x = 2$ makes the denominator zero. Nevertheless, the expansion tells us that the “badness” at 2 is not too bad. Specifically, we can see that while $g(2)$ is undefined, $(x-2)g(2)$ makes sense and is equal to $-\frac{2}{3}$.

It is easy to keep going. To handle general functions locally at a , we may sometimes need to use fractional powers of $(x-a)$, but it does not get much worse than that. Such expansions are a very powerful tool in the theory of functions. One of the motivations for the discovery of the p -adic numbers was to find a similarly powerful tool for the study of numbers.

2 Numbers Are Like Functions

It was DEDEKIND [VI.50] and Heinrich Weber who first realized that an analogy could be drawn between numbers and functions. In their scheme, positive whole

numbers were compared to polynomials, while fractions were analogous to quotients of polynomials such as the function $g(x)$ above. More complicated functions were like more complicated kinds of number. ELLIPTIC FUNCTIONS [V.34], for example, were similar to certain kinds of algebraic number. On the other hand, functions like $\sin(x)$ were more like TRANSCENDENTAL NUMBERS [III.43] such as e or π .

Dedekind and Weber pushed the idea that “functions are like numbers” in order to understand functions better. In particular, they showed that the techniques developed to study algebraic numbers could be used to study a whole class of functions, which came to be known as algebraic functions. It was Kurt Hensel, however, who saw that if functions are like numbers, then numbers must be like functions. In particular, he set out to find an analogue, for numbers, of the local expansions that were so useful in the theory of functions.

To get to Hensel’s idea, let us start by noticing that the way we usually represent numbers already points in the right direction. After all, an expression like 34 291 really means

$$34\,291 = 1 + 9 \cdot 10 + 2 \cdot 10^2 + 4 \cdot 10^3 + 3 \cdot 10^4 + 3 \cdot 10^5.$$

If we allow ourselves to think of 10 as being something like the variable x , this looks exactly like a polynomial. What is more, just as we can expand a polynomial in terms of different expressions $(x-a)$, we can write numbers in other bases. For example,

$$34\,291 = 4 + 4 \cdot 11 + 8 \cdot 11^2 + 3 \cdot 11^3 + 2 \cdot 11^4.$$

It is easy to see how to find this expansion. First, divide 34 291 by 11, and look at the remainder. It is 4. That is our first term. Next, subtract 4 from the original number to get something divisible by 11:

$$34\,291 - 4 = 34\,287 = 3117 \cdot 11.$$

Now divide 3117 by 11 to find the next remainder, which will give the second term. Keep repeating this process, and you will find the base-11 expansion.

That sounds very promising, but there is one little insight missing. The fact is that 10 is not really like $(x-2)$, because 10 *can be factored*, while $(x-2)$ cannot. So expanding a number in base 10 is a little like trying to express a polynomial in powers of $(x^2 - 3x + 2)$, which factors as $(x-1)(x-2)$. Such an expansion is not really local, since it is looking at two possible values of x at once. Similarly, the base-10 expansion mixes information about 2 and information about 5. The upshot is that we should always use a *prime number* as our base.

Just to fix ideas, let us choose $p = 11$. We already know that we can write positive numbers in base 11, i.e., as “polynomials in powers of 11.” What happens if we try it with a fraction? Let us take $\frac{1}{2}$. The first step is to find the remainder, that is, to find a number r (positive, between 0 and 10) such that $\frac{1}{2} - r$ is divisible by 11. Well, $\frac{1}{2} - 6 = -\frac{11}{2} = -\frac{1}{2} \cdot 11$. So the first term is 6. (To see what is meant by divisibility here, consider what would have happened if we had taken $r = 4$. Then $\frac{1}{2} - r$ would have been $-\frac{7}{2}$, and if we divide that by 11 we get $-\frac{7}{22}$, which has a factor of 11 in the denominator. It is this that is not allowed and that does not happen when $r = 6$.)

Now we repeat with the quotient, which was $-\frac{1}{2}$. We see that $-\frac{1}{2} - 5 = -\frac{11}{2} = -\frac{1}{2} \cdot 11$. So the second term will be $5 \cdot 11$. But now we find ourselves having to do $-\frac{1}{2}$ again! So we will do this again and again, and *all* of the remaining terms will have coefficient 5. In other words,

$$\frac{1}{2} = 6 + 5 \cdot 11 + 5 \cdot 11^2 + 5 \cdot 11^3 + 5 \cdot 11^4 + 5 \cdot 11^5 + \cdots$$

It is not clear quite what the equals sign means here, but in any case we have obtained an infinite expansion in powers of 11. It is called the *11-adic expansion* of $\frac{1}{2}$. Furthermore, the expansion “works” when we do arithmetic with it. For example, if we multiply it by 2 and do all the rearranging ($2 \times 6 = 12 = 1 + 11$, so carry a 1, etc.) we do end up with 1.

Hensel showed that one can do this with all algebraic numbers as long as one allows infinite expansions, a finite number of negative powers of 11 (so that one can handle $\frac{5}{33}$ and similar things), and, in certain cases, fractional powers of 11. He argued that we should view such expansions as giving information “locally at 11.” The same happens with all of the prime numbers. So if we have a prime number p we can consider our numbers “locally at p ” by taking their expansions in powers of p . These we call their *p-adic expansions*. Just as in the case of functions, such expansions immediately tell us how divisible by p a number is, while hiding all the information about other primes; in that sense, they are truly “local.”

3 p-adic Numbers

The best answers always raise new questions. Having discovered that any rational number has a p -adic expansion, and that one can “do arithmetic” directly with the expansions, it is inevitable to ask whether we have therefore enlarged the world of numbers under consideration. Once we have chosen the prime p , any

rational number gives us a p -adic expansion. But does every such expansion come from a rational number?

Not a chance. It is easy to see that the set of all expansions is much bigger than the set of all rational numbers. Hensel’s next move, then, was to point out that the set \mathbb{Q}_p of all possible p -adic expansions is a new realm of numbers, which he called the *p-adic numbers*. It includes not only all the rational numbers, but also a lot more.

The best way to think of \mathbb{Q}_p is by analogy with the set \mathbb{R} of all real numbers. Real numbers are usually given by their decimal expansions. When we write $e = 2.718\dots$, what we mean is that

$$e = 2 + 7 \cdot 10^{-1} + 1 \cdot 10^{-2} + 8 \cdot 10^{-3} + \cdots$$

The set of all such expansions is the set of all real numbers. It contains all the rational numbers, but is much bigger.

Of course, except for the fact that both contain the rationals, these two realms are almost completely different. For example, in both \mathbb{Q}_p and \mathbb{R} there is a natural notion of “distance between two numbers.” But these distances are completely different, even when the numbers in question are rational. So, in the reals, 2 is very close to 2001/1000. In the 5-adics, however, the distance between these two numbers is quite large!

It turns out that we can do calculus with p -adic numbers, just as we do it with reals. Many other mathematical ideas also extend. So Hensel’s ideas led to a system of “parallel (numerical) universes”—one for each prime, plus the real numbers—in which we can do mathematics.

4 The Local-Global Principle

At first, most mathematicians seem to have found Hensel’s new numbers interesting in a formal way, but also to have wondered what the point of them was. One does not adopt a new number system just for fun; it needs to be useful for something. Hensel was fascinated by his numbers and kept writing about them, but to begin with he had trouble demonstrating their usefulness. He showed, for example, that they could be used to develop the basics of algebraic number theory in a new way—but most folks seemed happy with the old way.

One can demonstrate the power of a new idea by giving a beautiful and easy proof of a difficult result. Hensel wrote a paper purporting to do just that: he gave an easy and elegant p -adic proof that the number e is transcendental. This did get people’s attention.

Unfortunately, when they looked hard at the proof they realized that it contained a subtle error. As a result, mathematicians' attitude of suspicion about Hensel's strange new numbers was reinforced.

The tide was turned by Helmut Hasse. He had been studying in Göttingen. At one point, he walked into a used bookstore and found a copy of Hensel (1913), a book written a few years earlier. Hasse was fascinated, and moved to Marburg to study with Hensel. A couple of years later, in 1920, he found the idea that was to make the p -adic numbers a crucial tool for number theorists.

What Hasse showed was that it was possible to answer some questions in number theory by answering them "locally." Here is a (not very important, but fairly easy to follow) example. Suppose x is a rational number that is a square of some other rational number y , so $x = y^2$. Since all rational numbers are also p -adic, it is true that *for every prime number p the number x , thought of as a p -adic number, is a square*. And similarly, the real number x is a square. In other words, the rational number y is a kind of "global" square root, in that it serves as a square root in each local setting.

So far, so boring. But now reverse the thing. Suppose that we know that *for every prime number p the number x , thought of as a p -adic number, is the square of some p -adic number (which may depend on p)*, and also that x , thought of as a real number, is the square of some real number. A priori, these local square roots of x could all be different! But it turns out that under these assumptions x must be the square of some *rational* number, so that in fact all the local roots must come from a "global" root.

This leads us to think of the rational numbers as "global" and of the various \mathbb{Q}_p and of \mathbb{R} as "local." Then the previous paragraph claims that the property of "being a square" is true globally if and only if it is true "everywhere locally." This turns out to be a powerful and illuminating idea, and it has become known as the *Hasse principle* or the *local-global principle*.

Our example, of course, demonstrates the principle in its strongest case: solve a problem locally in all cases, and you have solved it globally. That is often too much to hope for. Nevertheless, attacking a problem locally and then putting the local pieces together has become a fundamental technique in modern number theory. It has been used to simplify older proofs, as in CLASS FIELD THEORY [V.30], and also to obtain new results, as in Wiles's proof of FERMAT'S LAST THEOREM [V.12]. So Hensel was right after all: his new numbers have

earned their place along with the real numbers in every number theorist's heart.

Further Reading

- Gouvêa, F. Q. 2003. *p -adic Numbers: An Introduction*, revised 3rd printing of the 2nd edn. New York: Springer.
- Hasse, H. 1962. Kurt Hensels entscheidener Anstoss zur Entdeckung des Lokal-Global-Prinzips. *Journal für die reine und angewandte Mathematik* 209:3–4.
- Hensel, K. 1913. *Zahlentheorie*. Leipzig: G. J. Göschenische.
- Roquette, P. 2002. History of valuation theory. I. In *Valuation Theory and Its Applications*, volume I, pp. 291–355. Providence, RI: American Mathematical Society.
- Ullrich, P. 1995. On the origins of p -adic analysis. *Proceedings of the 2nd Gauss Symposium. Conference A: Mathematics and Theoretical Physics, Munich, 1993*, pp. 459–473. Symposia Gaussiana. Berlin: Walter de Gruyter.
- . 1998. The genesis of Hensel's p -adic numbers. In *Charlemagne and His Heritage. 1200 Years of Civilization and Science in Europe*, volume 2, pp. 163–78. Turnhout: Brepols.

The Logarithmic Function

See THE EXPONENTIAL AND LOGARITHMIC FUNCTIONS [III.25]

III.54 The Mandelbrot Set

Suppose we have a complex polynomial f defined by a formula $f(z) = z^2 + C$ for some complex number C . Then for any choice of complex number z_0 we can form a sequence z_0, z_1, z_2, \dots by *iterating*, that is, repeatedly applying, the function f . So we let $z_1 = f(z_0)$, $z_2 = f(z_1)$, and so on. Sometimes the resulting sequence will tend to infinity, but sometimes it remains bounded—that is, it stays within a fixed distance from 0. For example, if we take $C = 2$ and start with $z_0 = 1$, then the sequence goes 1, 3, 11, 123, 15 131, ... and clearly tends to infinity, whereas if we start with $z_0 = \frac{1}{2}(1 - i\sqrt{6})$, then we find that $z_1 = z_0^2 + 2 = z_0$ so the sequence is bounded since all its terms are equal to z_0 . The *Julia set* associated with the constant C is the set of all z_0 for which the sequence remains bounded. Julia sets often have a fractal shape (see [IV.15 §2.5]).

To define a Julia set, one fixes C and considers different possibilities for z_0 . What happens if one fixes z_0 and considers different possibilities for C ? The result is the *Mandelbrot set*. The precise definition is that it is the set of all C such that the sequence is bounded if you take $z_0 = 0$. (One could consider other values

of z_0 , but the resulting sets are not interestingly different because they are related to each other by a simple change of variables.)

The Mandelbrot set also has an intricate fractal shape—one that has captured the popular imagination. The detailed geometry of the Mandelbrot set is not yet fully understood; some of the resulting open problems are of major importance because they encode very general information about dynamical systems. See DYNAMICS [IV.15 §2.8] for more details.

III.55 Manifolds

The surface of a sphere has the property that if you look at a very small portion of it then that portion will look like part of a plane. More generally, a d -dimensional manifold, or d -manifold, is a geometrical object that looks “locally” like d -dimensional EUCLIDEAN SPACE [I.3 §6.2]. Thus, 2-manifolds are smooth surfaces such as those of a sphere or a torus. Higher-dimensional manifolds are harder to visualize, but are a major topic of research. The basics of manifolds are set out in SOME FUNDAMENTAL MATHEMATICAL DEFINITIONS [I.3 §§6.9, 6.10]. More advanced ideas are discussed in DIFFERENTIAL TOPOLOGY [IV.9] and ALGEBRAIC TOPOLOGY [IV.10]. See also ALGEBRAIC GEOMETRY [IV.7], MODULI SPACES [IV.8], and RICCI FLOW [III.80]. (Even this is far from a complete list of articles in which manifolds feature.)

III.56 Matroids

Dominic Welsh

The original aim of Hassler Whitney when he introduced the concept of a matroid in 1935 was to produce an abstract notion that would capture the main ingredients of the structure of a set of vectors in a VECTOR SPACE [I.3 §2.3], while avoiding any explicit mention of linear independence.

To do this he singled out two fundamental properties and postulated that any family of subsets that possessed these properties was the collection of “independent sets” of a “matroid.” The first of these properties was an obvious one: any subset of a linearly independent set is also linearly independent. The second property was more subtle: if A and B are two linearly independent sets and B contains more elements than A , then there exists some element of B that is not in A but which, when added to A , gives a set that is still linearly independent. Finally, in order to avoid trivialities

he insisted that in every matroid the empty set must be independent.

Thus, formally, a *matroid* is defined to be a finite set E together with a family of subsets of E which are called the *independent sets* and which satisfy the following axioms.

- (i) The empty set is independent.
- (ii) Every subset of an independent set is independent.
- (iii) If A and B are independent sets, with the number of elements of A being one less than the number of elements of B , then there is some x in B that is not in A such that $A \cup \{x\}$ is also independent.

Property (iii) is called the *exchange axiom*. The most fundamental example of a matroid is a set of vectors in a vector space with the “independent sets” being the usual linearly independent ones: in this case the exchange axiom is known as Steinitz’s exchange lemma. However, there are many examples of matroids that are not subsets of vector spaces.

Here, for example, is an important class of matroids that arise from graph theory. A *cycle* in a graph is a collection of edges of the form $(v_1, v_2), (v_2, v_3), \dots, (v_{k-1}, v_k), (v_k, v_1)$, where the v_i are distinct vertices. Take any graph and call a subset of edges “independent” if it contains no cycle.

So here we are thinking of a cycle among the edges as being in some way similar to a linear dependence among some vectors. It is obvious that any subset of an independent set will also not contain a cycle, so condition (ii) is satisfied. Slightly less obvious is that if A and B are sets of t and $t + 1$ edges, respectively, neither containing a cycle, then there will be at least one edge in B but not in A which can be added to A without creating a cycle. So we see that this is another example of a matroid, even though it arises in a very different context from the vector space one.

As it turns out, there is a way of identifying the edges of a graph with a set of vectors in a vector space over the field \mathbb{F}_2 of integers mod 2 (see MODULAR ARITHMETIC [III.60]). If G has n vertices and one associates with each vertex a basis element of \mathbb{F}_2^n , then one can associate with each edge the vector that is given by the sum of the basis elements corresponding to its two endpoints. A set of edges will then be independent if and only if the corresponding vectors in \mathbb{F}_2^n are linearly independent. However, as we shall see, there are important examples of matroids that are not even *isomorphic* to sets of vectors.

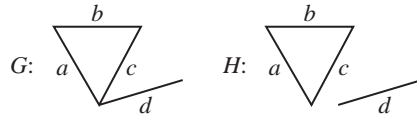


Figure 1 Two graphs giving rise to the same matroid.

Note that the collection of the independent sets (in a graph) conveys part of the information present in the graph, but by no means all of it. For example, consider the graphs G and H in figure 1. As graphs, G and H are distinct, but both give the same matroid on the set $\{a, b, c, d\}$ (the independent sets are all subsets of size less than or equal to 3, except for $\{a, b, c\}$). Note that this matroid is also the same as the matroid formed by the columns of the matrix

$$A = \begin{pmatrix} a & b & c & d \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

However, it turns out that most matroids do not come from either graphs or matrices.

Although a matroid is defined by very simple axioms, many basic results from both linear algebra and graph theory can be extended to the wider setting of matroids. For example, suppose that T is a connected graph. It is not hard to prove that if B is a maximal independent set of the matroid on G , then B is a tree which is incident with every vertex of G . Such a tree is called a *spanning tree* of G . All spanning trees of a connected graph have the same number of edges, namely, one less than the number of vertices. Similarly, in a vector space, or indeed in any subset of vectors, all maximal linearly independent sets have the same size. Both of these are special cases of the general result that in any matroid all maximal independent sets have the same size. This common size is called the *rank* of the matroid and, by analogy with vector spaces, a maximal independent set in a matroid is called a *base*.

Matroids arise naturally in many parts of mathematics, and they often turn up unexpectedly. For example, consider the *minimum connector problem*: a company needs to connect a number of cities by links, such as railways or phone cables, and wishes to minimize the total cost. This is clearly equivalent to the following problem. Given a connected graph G , with each edge e having a nonnegative weight $w(e)$, find a set of edges that has the minimum total weight but that connects all

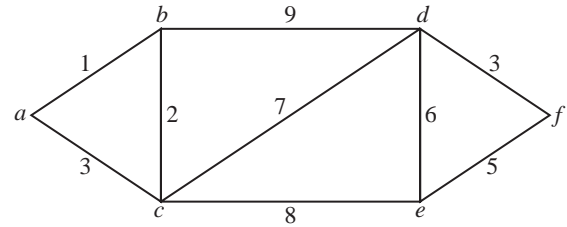


Figure 2 A graph with edge-weights.

the vertices of G . It is not hard to see that this problem reduces to finding a spanning tree of minimum weight.

For this there is a classical algorithm. It is the simplest possible algorithm one could imagine for the problem, and it works as follows. Start by choosing an edge of minimum weight, and at each subsequent step add an edge of minimum weight to your chosen set provided that at no stage a cycle is formed.

For example, consider the graph in figure 2. The algorithm would successively select the edges (a, b) , (b, c) , (d, f) , (e, f) , (c, d) , giving a spanning tree of total weight $1 + 2 + 3 + 5 + 7 = 18$. Because of the way it works, the algorithm is known as a *greedy algorithm*.

At first sight, it seems rather unlikely that this algorithm could work, as it denies the possibility that choosing a suboptimal edge now might have a payoff later. However, it is not hard to show that the algorithm is actually correct. In fact, it extends in almost exactly the same way to matroids in general: what it gives is a (rather fast) algorithm for selecting a base of minimum weight in a matroid in which each element has a nonnegative weight.

Somewhat more surprisingly, matroids are the only structures for which the greedy algorithm works. More precisely, suppose that \mathcal{I} is a family of subsets of a set E with the property that if $A \in \mathcal{I}$ and $B \subseteq A$, then $B \in \mathcal{I}$. Now let w be any weight function and suppose that the problem is to select a member B of \mathcal{I} which has maximum weight, where the weight of a set is just the sum of the weights of its elements. As above, the greedy algorithm starts by choosing an element e of maximum weight and then successively picks elements of maximum weight from the remaining elements subject to the proviso that at each stage, the set of elements chosen is a member of \mathcal{I} . It turns out that the following is true: *the greedy algorithm works on \mathcal{I} for all weight functions w if and only if \mathcal{I} is the collection of independent sets of a matroid*. Thus, matroids form a “natural home” for many optimization problems. Moreover, the concept is genuinely useful, since many of the matroids

PUP: I can confirm that the numbering of the lines in the figure is OK as it stands.

T&T note: change made here has to be ratified by Imre and/or the author at some point.

PUP: Tim definitely prefers ‘and’ here. OK?

PUP: Tim prefers ‘but’ here. OK?

that arise in such problems are not derived from either vector spaces or graphs.

III.57 Measures

To understand measure theory, and to see why it is useful and important, it is instructive to start with a problem about lengths. Suppose that we have a sequence of intervals in $[0, 1]$ (the closed interval from 0 to 1), of total length less than 1. Can they cover $[0, 1]$? In other words, given intervals $[a_1, b_1], [a_2, b_2], \dots$, with $\sum(b_n - a_n) < 1$, is it possible that their union equals $[0, 1]$?

One is tempted to answer “no, as the total length is too small.” But this is just to restate the question. After all, why should “total length less than 1” actually imply that the intervals cannot cover $[0, 1]$? Another tempting answer is to say “just rearrange the intervals so that they go from the left to the right, and then we never get to the right-hand end of $[0, 1]$.” In other words, if the n th interval has length $b_n - a_n = d_n$, then just translate the intervals to be the intervals $[0, d_1], [d_1, d_1 + d_2], \dots$. In this rearrangement, it is indeed true that we never cover any point beyond $\sum d_n$, and so do not cover all of $[0, 1]$, but why does this imply that the original intervals do not cover $[0, 1]$?

It is quite easy to see that this rearrangement argument works for a *finite* number of intervals, but it does not work in general. Indeed, suppose we ask the same question, but for the rationals: that is, let us replace the interval $[0, 1]$ by the rational interval $[0, 1] \cap \mathbb{Q}$. If our intervals have lengths $\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \dots$, for example, so that the total length is only $\frac{1}{2}$, then certainly the left-to-right intervals will cover only the interval $[0, \frac{1}{2}] \cap \mathbb{Q}$, but it is possible for the original intervals to cover all of $[0, 1] \cap \mathbb{Q}$, since we can just enumerate the rationals as q_1, q_2, \dots (see COUNTABLE AND UNCOUNTABLE SETS [III.11]), and then put an interval of length $\frac{1}{4}$ around q_1 , one of length $\frac{1}{8}$ around q_2 , and so on.

This observation shows that the answer to our problem must involve properties of the reals that are not shared by the rationals—which wrecks any kind of “it is obvious” argument. In fact, the result is true for the reals, but its proof is a good exercise.

Why is this an important fact? It stems from a wish to define “length” for general sets of reals (for simplicity, we will concentrate on $[0, 1]$, just to avoid some technicalities about “infinite length”). What should the “length” of a set be? For intervals the answer is clear,

and it is also clear for finite unions of intervals. But what about sets like $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots\}$, or \mathbb{Q} itself?

A natural first attempt would be to use finite unions of intervals: one could take the length of a set A to be the least value of the length of a finite union of intervals that covers A . More precisely, one could define the length of A to be the infimum of $(b_1 - a_1) + \dots + (b_n - a_n)$, taken over all finite unions of intervals $[a_1, b_1] \cup \dots \cup [a_n, b_n]$ that cover A .

Unfortunately, this definition has some very undesirable properties. For example, the length of the set of all rational numbers in the interval $[0, 1]$ would then be 1, as would the length of all irrational numbers in $[0, 1]$. We would thus have two disjoint sets (and very natural ones at that) such that the length of their union is not the sum of their lengths. So this form of “length” is not really well-behaved for such sets.

What we want is a notion of length that applies to all the sets we know and are used to, and is *additive*, meaning that the length of $A \cup B$ is the sum of the lengths of A and B whenever A and B are disjoint. Remarkably, this *can* be achieved, and the key idea is to allow *countable* covers. That is, we modify the above definition as follows: the length (or *measure*, to give it its usual name) of a set A is the infimum of $(b_1 - a_1) + (b_2 - a_2) + \dots$, taken over all unions of intervals $[a_1, b_1] \cup [a_2, b_2] \cup \dots$ that cover A . Note that, thanks to the puzzle discussed earlier, the measure of the interval $[a, b]$ is $b - a$, just as we would hope.

It is also not hard to see that the measure of the set of rationals in $[0, 1]$ is zero, and it turns out that the measure of the irrationals in $[0, 1]$ is 1. Indeed, any countable set has measure zero. In many contexts, sets of measure zero are regarded as “negligible” or “of no importance.” It is worth mentioning that there are also sets of measure zero that are uncountable (an example is the CANTOR SET [III.17]).

It turns out that, even with this definition, there are pairs of disjoint sets A and B such that the measure of $A \cup B$ is not the sum of the measures of A and B . However, it can be shown that for all “reasonable” sets the measure is additive. More precisely, one says that a subset of $[0, 1]$ is *measurable* if the measures of it and its complement add up to 1, as they should. If A and B are disjoint measurable sets, then the measure of their union is the sum of their measures.

This is a very useful fact, since it can be shown that every set that arises naturally in mathematics, or that has an explicit definition, is measurable: intervals, finite

unions of intervals, countable unions of intervals, Cantor sets, things involving rationals or irrationals, and so on. In fact, the union of any countable family of measurable sets is again measurable (one says that the measurable sets form a *sigma-algebra*). Even better, for measurable sets the measure is *countably* additive, meaning that the measure of a disjoint union of countably many measurable sets is the sum of the measures of the individual sets.

More generally, in many other settings, one wants to end up with a sigma-algebra, containing all the sets one is interested in, on which we can define a countably additive measure, or “length function.” The above example is called *Lebesgue measure on $[0, 1]$* . In general, whenever one wishes to define a countably additive measure, one always needs a result like the puzzle above in order to get started.

Here is another example: we could work in $[0, 1]^2$ (the unit square in the plane), and base our ideas upon rectangles instead of intervals. So we would define the measure of a set as the least total area of a sequence of rectangles that covers the set. This gives an elegant and powerful approach to integration: the integral of a function f (defined on $[0, 1]$, say, and taking values in $[0, 1]$) is just defined to be the “area under its graph”: that is, the measure of the set $\{(x, y) : y \leq f(x)\}$. Many complicated-looking functions can now be integrated: for example, the function f that is 1 on the rationals and 0 on the irrationals is easily checked to have an integral, namely 0, whereas in earlier theories such as Riemann integration that function would be too rapidly varying to be integrable.

This approach to integration gives rise to the so-called *Lebesgue integral* (further discussed in the article on LEBESGUE [VI.72]), which is one of the fundamental concepts in mathematics. It allows one to integrate a wide range of functions that are not Riemann integrable, but the main reason for its importance is not so much this as the fact that the Lebesgue integral has very good limiting properties that the Riemann integral lacks. For example, if f_1, f_2, \dots is a sequence of Lebesgue-integrable functions from $[0, 1]$ to $[0, 1]$ and $f_n(x)$ converges to $f(x)$ for every x , then f is Lebesgue-integrable and the Lebesgue integrals of the functions f_n converge to the Lebesgue integral of f .

III.58 Metric Spaces

There are many contexts in mathematics, especially in analysis, where one would like to say that two

mathematical objects are close, and understand precisely what that means. If the two objects are the points (x_1, x_2) and (y_1, y_2) in a plane, then the task is straightforward: the distance between them is $\sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2}$, by the Pythagorean theorem, and it makes sense to say that the points are close if this distance is small.

Now suppose that we have two points in n -dimensional space, (x_1, \dots, x_n) and (y_1, \dots, y_n) . It is a simple matter to generalize the formula just given when $n = 2$ and define the distance between them to be

$$\sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2 + \dots + (y_n - x_n)^2}.$$

Of course, the fact that the formula can be easily generalized is not in itself a guarantee that the resulting notion is a sensible definition of distance. And this raises the question of what properties we would like a definition to have for it to count as sensible. A metric space is an abstract notion that results from thinking about this question.

Let X be a set of “points.” Suppose that, given any two of these points, x and y say, we have a way of assigning a real number $d(x, y)$ that we wish to regard as the distance between them. The following three properties are ones that it would be highly desirable for this idea of distance to have.

- (P1) $d(x, y) \geq 0$ with equality if and only if $x = y$.
- (P2) $d(x, y) = d(y, x)$ for any two points x and y .
- (P3) $d(x, y) + d(y, z) \geq d(x, z)$ for any three points x , y , and z .

PUP: list items changed to (P1), (P2), (P3) here and (i), (ii), (iii) below – OK?

The first of these properties says that the distance between two points is always positive, except when the two points are the same, when it is zero. The second says that distance is a *symmetric* notion: the distance from x to y is the same as the distance from y to x . The third is called the *triangle inequality*: if you imagine x , y , and z as the vertices of a triangle, it says that the length of any side never exceeds the sum of the lengths of the other two sides.

A function d defined on pairs of points (x, y) from a set X is called a *metric* if it has properties (P1)–(P3) above. In that case, X and d together form a *metric space*. This abstraction of the usual notion of distance is very useful, and there are many important examples of metrics that are not necessarily derived from the Pythagorean theorem. Here are a few examples.

- (i) Let X be n -dimensional space, that is, the set \mathbb{R}^n of all sequences (x_1, \dots, x_n) of n real numbers.

It can be shown that the formula derived above from the Pythagorean theorem gives a notion of distance that does indeed satisfy properties (P1)–(P3). This metric is called the *Euclidean distance* and the resulting metric space is called *Euclidean space*. Euclidean spaces are perhaps the single most basic and important class of metric spaces in mathematics.

- (ii) Information is often transmitted digitally in the form of a string of 0s and 1s, such as 000111010010. The *Hamming distance* between two such strings is defined to be the number of places where the strings are different. For example, the Hamming distance between the strings 00110100 and 00100101 is 2, since the strings differ in the fourth and eighth places only. This idea of distance also satisfies properties (P1)–(P3).
- (iii) If you are driving from one town to another, then the distance you care about is not the distance as the crow flies but the length of the shortest route along the network of available roads. Similarly, if you wish to travel from London to Sydney, then what matters is the length of the shortest path (known as a *geodesic*) along the Earth's surface, rather than the “actual” distance through the Earth itself. Many useful metrics come from this general idea of a shortest route, which guarantees that property (P3) will hold.
- (iv) An important feature of Euclidean distance is its rotational symmetry: in other words, rotating the plane, or space, does not alter the Euclidean distances between points. There are other metrics that also have a great deal of symmetry, and these have great geometrical significance. In particular, the discovery of the *HYPERBOLIC METRIC* [I.3 §§6.6, 6.10] in the early nineteenth century demonstrated that the parallel postulate could not be proved using Euclid's other axioms. This resolved a question that had been open for thousands of years. See *RIEMANNIAN METRICS* [I.3 §6.10].

III.59 Models of Set Theory

A model of set theory is, roughly speaking, a structure in which the usual axioms of set theory (ZF, or ZFC) hold. To explain what this means, let us think first about groups. The axioms of group theory mention certain operations (such as multiplication and inversion),

and a model of group theory is a set, equipped with such operations, such that the axioms hold. In other words, a model of group theory is nothing other than a group. So what does a “model of ZF” mean? The axioms of ZF mention one relation, namely “is an element of,” or “ \in .” A model of ZF is a set M , on which there is a relation E , such that all the axioms of ZF hold in S if we replace “ \in ” by “ E .”

However, there is one very important difference between these two sorts of model. When one first meets groups, one starts with some very simple examples, such as cyclic groups, or groups of symmetries of regular polygons, and one then builds up to more sophisticated examples such as the *SYMMETRIC AND ALTERNATING GROUPS* [III.70], and beyond. But this gentle process is not available for models of ZF. Indeed, since all of mathematics can be formulated in the language of ZF, it follows that *every* model of ZF has to contain a “copy” of the whole world of mathematics. This makes studying models of ZF rather difficult.

One aspect that is often found puzzling is the fact that a model of ZF is a *set*. This might seem to mean that there is a “universal” set (a set that has every set as a member), but from *RUSSELL'S PARADOX* [II.7 §2.1] it is easy to see that there can be no such set. The answer to this apparent problem is that the model M is indeed a set in the real mathematical universe, but that inside the model there is no universal set—in other words, there is no element x of M such that $\forall y Ex$ for every element y of M . Thus, from the perspective of the model, the statement “there is no universal set” is true.

See *MODEL THEORY* [IV.2] for more about models in general, and *SET THEORY* [IV.1] for more about models of set theory.

III.60 Modular Arithmetic

Ben Green

Is there a square number whose decimal expansion ends ...7? Is 438345 divisible by 9? For which positive integers n is $n^2 - 5$ a power of two? Is $n^7 - 77$ ever a Fibonacci number?

These questions, and more, can be answered using modular arithmetic. Let us look at the first question. Listing the first few squares, 1, 4, 9, 16, ..., one does not find any whose final digit is 7. In fact, writing down just the final digits, one gets the sequence

1, 4, 9, 6, 5, 6, 9, 4, 1, 0, 1, 4, 9, 6, 5, 6, ...

PUP: thanks to the proofreader for spotting strange mistake here!

which seems to repeat (and thus never contain the number 7).

An explanation of this phenomenon is as follows. Let n be a number to be squared. We can always write n as a multiple of 10 plus a remainder; that is, $n = 10q + r$, where $r \in \{0, 1, \dots, 9\}$. Now, if we square n we get

$$\begin{aligned} n^2 &= (10q + r)^2 \\ &= 100q^2 + 20qr + r^2 \\ &= 10(10q^2 + 2r) + r^2. \end{aligned}$$

The only part of this expression that affects the final digit is the r^2 , which immediately explains why the sequence of last digits of squares repeats with period 10, and hence contains no 7s.

Modular arithmetic is essentially just a notation for writing down arguments of this sort. If two numbers (like n and r) leave the same remainder on division by 10, then we say that they are *congruent modulo* 10 and write $n \equiv r \pmod{10}$. What we proved above is the statement that, if $n \equiv r \pmod{10}$, then $n^2 \equiv r^2 \pmod{10}$.

Everything we have just said applies equally well if we replace 10 by an arbitrary *modulus* m : if n and r leave the same remainder on division by m , then we say that n and r are *congruent modulo* m and we write $n \equiv r \pmod{m}$. Equivalently, n and r are congruent modulo m if m divides $n - r$. (An integer a is said to *divide* another integer b if b is an integer multiple of a .) The above argument is just one instance of the following general fact, which is not hard to prove: if $a \equiv a' \pmod{m}$ and $b \equiv b' \pmod{m}$, then $ab \equiv a'b' \pmod{m}$ and $a + b \equiv a' + b' \pmod{m}$.

Now observe that $10 \equiv 1 \pmod{9}$. It follows that $10 \times 10 \equiv 1 \times 1 \equiv 1 \pmod{9}$, and in fact that $10^d \equiv 1 \pmod{9}$ for any $d \in \mathbb{N}$. Suppose that we have a number N whose decimal expansion is $a_d a_{d-1} \dots a_2 a_1 a_0$. This means that

$$N = a_d 10^d + a_{d-1} 10^{d-1} + \dots + a_1 10 + a_0.$$

Applying the rules of modular arithmetic, we get

$$N \equiv a_d + \dots + a_{d-1} + \dots + a_1 + a_0 \pmod{9}.$$

This gives the well-known test for divisibility by 9: simply add up the digits of the number in base 10, and see if the result is divisible by 9. For the example $N = 438345$ the sum of the digits is 27, which is divisible by 9. So N is a multiple of 9 (in fact $N = 9 \times 48905$).

If m is a modulus and n is an integer, then there is precisely one value of r between 0 and $m - 1$ such that $n \equiv r \pmod{m}$. This number r is often called the least residue or simply the *residue* of n to the modulus m .

Now let us consider the third question posed at the beginning of this article, namely the matter of when $n^2 - 5$ is a power of two. When $n = 3$, $3^2 - 5 = 4$ is a power of two, but a little experimentation does not reveal any further examples. What aspect of the problem changes as n becomes larger than 3? The key observation is that $n^2 - 5$ is now greater than 4, and so if it were a power of 2, then it would have to be divisible by 8. That would mean that $n^2 \equiv 5 \pmod{8}$, but this is never the case. Indeed, the residues of the first eight squares are 1, 4, 1, 0, 1, 4, 1, 0, and we know that the sequence will repeat with period 8 (actually, it repeats with period 4). Thus it never contains a 5.

Modular arithmetic should be used with care. Although the rules for addition and subtraction are simple, division is somewhat more subtle. For example, if one has an equation $ac \equiv bc \pmod{m}$, it is not, in general, permissible to divide by c and conclude that $a \equiv b \pmod{m}$: consider, for instance, the case $a = 2$, $b = 4$, $c = 3$, $m = 6$.

Let us examine what has just gone wrong. To say that $ac \equiv bc \pmod{m}$ means that m divides $ac - bc = (a - b) \times c$. But this clearly does not mean that m divides $a - b$, since m could divide c (or at least have a common factor with it). However, if m has no factor in common with c , then it must divide $a - b$, so in this case we do indeed have $a \equiv b \pmod{m}$. In particular, for any prime number p we have the very useful *cancellation law*: if $ac \equiv bc \pmod{p}$ and $c \not\equiv 0 \pmod{p}$, then $a \equiv b \pmod{p}$.

The examples so far may have suggested that the principal uses of modular arithmetic are to do with specific moduli such as 10 and 8. However, this is far from true, and the subject really comes into its own when one looks at more general m . For example, one of the basic results in number theory is *Fermat's little theorem*, which states that if p is a prime and $a \not\equiv 0 \pmod{p}$, then $a^{p-1} \equiv 1 \pmod{p}$. Let us quickly prove this. Consider the numbers $a, 2a, 3a, \dots, (p-1)a \pmod{p}$. If $ra \equiv sa \pmod{p}$, then from the cancellation law we can deduce that $r \equiv s \pmod{p}$, from which it follows that $a, 2a, \dots, (p-1)a$ are all different modulo p . Furthermore, none of these numbers is $0 \pmod{p}$. We are thus forced to conclude that the numbers $a, 2a, 3a, \dots, (p-1)a \pmod{p}$ are simply a rearrangement of the numbers $1, 2, 3, \dots, p-1 \pmod{p}$. In particular, the products of the numbers in these two sets are the same, which implies that

$$a^{p-1} (p-1)! \equiv (p-1)! \pmod{p}.$$

Since $(p-1)!$ is not a multiple of p , we can apply the cancelation law again and divide both sides by $(p-1)!$. This implies the result.

Euler's theorem is a generalization of Fermat's little theorem to composite moduli. It states that if m is a positive integer and a is another positive integer that is *coprime* to m (this means that a and m have no common factor), then $a^{\phi(m)} \equiv 1 \pmod{m}$. Here ϕ is *Euler's totient function*: $\phi(m)$ is the number of integers less than m that are coprime to m . For instance, if $m = 9$, then the integers less than m and coprime to m are 1, 2, 4, 5, 7, and 8, so $\phi(9) = 6$ and we can deduce from Euler's theorem that $5^6 \equiv 1 \pmod{9}$. Let us check this directly: $5^6 = 15\,625$, so the sum of its digits is 19, which is indeed congruent to 1 mod 9. For further discussion of the Fermat-Euler theorem, see MATHEMATICS AND CRYPTOGRAPHY [VII.7], COMPUTATIONAL NUMBER THEORY [IV.5], and THE WEIL CONJECTURES [V.38].

The final question from above—whether $n^7 - 77$ is ever a Fibonacci number—is left as an exercise to the reader.

III.61 Modular Forms

Kevin Buzzard

1 A Lattice in the Complex Numbers

When one first learns about the complex numbers, one is taught to think of them as a two-dimensional space, with one real and one imaginary dimension: a complex number $z = x + iy$ has real part x and imaginary part y , where i is a square root of -1 .

Now let us consider what the complex numbers that have *integers* for their real and imaginary parts look like. These complex numbers, such as $3 + 4i$ or $-23i$, form a “lattice” in the complex plane (see figure 1).

By definition, every element of this lattice is of the form $m + ni$ for some pair of integers m and n . We say that the lattice is *generated* by 1 and i , and use the notation $\mathbb{Z} + \mathbb{Z}i$ for it. Note that this lattice can be generated in plenty of other ways. For example, it is also generated by the pair $(1, -i)$, the pair $(1, 100 + i)$ or even the pair $(101 + i, 100 + i)$. In fact, one can easily check that this lattice is generated by the pair $(a + bi, c + di)$ (meaning that every element of the lattice is an integer combination of $a + bi$ and $c + di$) if and only if a, b, c , and d are integers and $ad - bc = \pm 1$.

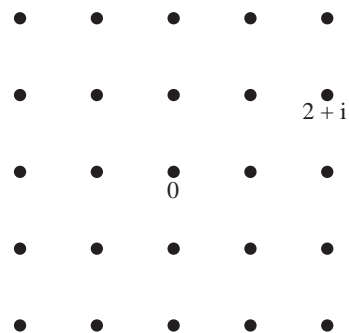


Figure 1 A lattice.

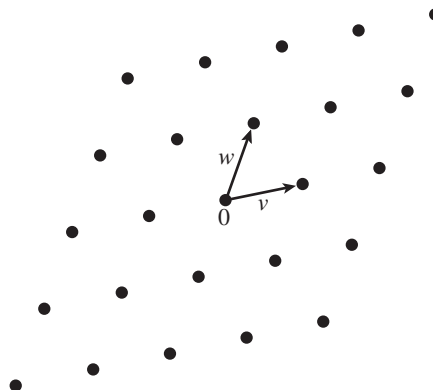


Figure 2 A general lattice.

2 More General Lattices

Now let v and w be *any* two complex numbers and consider the set of complex numbers of the form $av + bw$, again with a and b integers (see figure 2).

A lattice is exactly such a thing: a grid $\mathbb{Z}v + \mathbb{Z}w$ in the complex plane generated by two complex numbers v and w , with the provisos that neither v nor w is zero and that v/w is not real (this is just to ensure that v and w do not both lie on one line).

If $\tau = x + iy$ is a complex number with $y \neq 0$, then there is a standard lattice associated with τ , namely $\mathbb{Z}\tau + \mathbb{Z}$. We call this lattice Λ_τ and note that $\Lambda_\tau = \Lambda_{-\tau}$. In general, however, distinct complex numbers τ give rise to distinct lattices—and furthermore there are plenty of lattices that are not equal to Λ_τ for any τ , for the simple reason that 1 belongs to Λ_τ for every τ .

3 Relations between Lattices

If Λ is a lattice generated by v and w , and α is a nonzero complex number, then one can multiply the entire situ-

ation by α and deduce that $\alpha\Lambda$ is the lattice generated by αv and αw . Geometrically, this says that one can rotate and rescale lattices.

If Λ is a lattice generated by v and w , and we scale it by dividing everything by w , then we get a new lattice $(1/w)\Lambda$, which is generated by v/w and $w/w = 1$. In particular, this new lattice is equal to Λ_τ for the complex number $\tau = v/w$.

It may seem like an odd thing to do, but one can apply this scaling trick to Λ_τ itself. The lattice Λ_τ is generated by $(\tau, 1)$ but also by any pair $(v, w) = (a\tau + b, c\tau + d)$, if a, b, c , and d are integers such that $ad - bc = \pm 1$. If we divide by $c\tau + d$ and set $\sigma = (a\tau + b)/(c\tau + d)$, then we see that

$$\frac{1}{c\tau + d}\Lambda_\tau = \Lambda_\sigma. \quad (1)$$

4 Modular Forms as Functions on Lattices

The formal definition of a modular form is rather unenlightening: it is a function that obeys certain boundedness conditions and transformation properties. One way of seeing where the transformation properties come from is to think about lattices. If k is an integer, then a *modular form of weight k* is a function f that associates a complex number $f(\Lambda)$ with any lattice Λ , and has the property that

$$f(\alpha\Lambda) = \alpha^{-k}f(\Lambda). \quad (2)$$

The function also has to satisfy some other properties (a differentiability condition and a boundedness condition), but the crucial property is the one above. If k is even and at least 4, then an example of a modular form of weight k is the *Eisenstein series* G_k defined by the formula

$$G_k(\Lambda) = \sum_{0 \neq \lambda \in \Lambda} \lambda^{-k}.$$

The assumption that k is at least 4 guarantees that the sum converges, and the evenness of k ensures that the function is nonzero.

We have seen that any lattice can be scaled so that it takes the form Λ_τ for some τ , so (2) implies that a modular form will be determined by its values on such lattices. If \mathcal{H} denotes the complex numbers with positive imaginary part, then, because $\Lambda_\tau = \Lambda_{-\tau}$, a modular form is in fact determined by its values on Λ_τ for $\tau \in \mathcal{H}$.

However, an arbitrary function on \mathcal{H} does not give us a modular form: equation (1) tells us that if f is a modular form and F is the function on \mathcal{H} defined by

$F(\tau) = f(\Lambda_\tau)$, then F must satisfy the equation

$$F\left(\frac{a\tau + b}{c\tau + d}\right) = (c\tau + d)^k F(\tau) \quad (3)$$

for every $a, b, c, d \in \mathbb{Z}$ such that $ad - bc = 1$. (The reason we exclude the case $ad - bc = -1$ is that $(a\tau + b)/(c\tau + d)$ would not be in the upper half-plane in this case.) This is the equation at the heart of the definition of a modular form.

Over the years, mathematicians have isolated other desirable properties that F should have in order to give a useful theory. Nowadays, modular forms are required to obey the additional properties that F is HOLOMORPHIC [I.3 §5.6] and that $F(x + iy)$ does not grow too quickly as y goes to $+\infty$; these assumptions imply that the vector space of weight k modular forms is finite dimensional. The Eisenstein series above do have these additional properties, and are the first basic examples of modular forms.

5 Why Modular Forms?

Modular forms have links with arithmetic, geometry, representation theory, and even physics. Modular forms also played a key role in the Taylor-Wiles proof of FERMAT'S LAST THEOREM [V.12]. Why is this? One general reason is that there are links between modular forms and other mathematical objects: here we briefly explain one of the links.

Lattices in the complex plane are related to ELLIPTIC CURVES [III.21]: the quotient of the complex numbers by a lattice is an elliptic curve, and every elliptic curve arises in this way. Hence to study elliptic curves, or families of elliptic curves, one can instead study families of lattices. One way of studying an object is by studying the functions on that object, and a modular form is *precisely* that: a function on the collection of all lattices. And indeed, automorphic forms, which are generalizations of modular forms, have been used to great effect in studying a wide variety of families of algebraic objects in this way.

III.62 Moduli Spaces

An important general problem in mathematics is *classification* (see THE GENERAL GOALS OF MATHEMATICAL RESEARCH [I.4 §2]). Often, one has a set of mathematical structures and a notion of equivalence, and one would like to describe the EQUIVALENCE CLASSES [I.2 §2.3]. For example, two (compact, orientable) surfaces are

often regarded as equivalent if each can be continuously deformed into the other. Each equivalence class is then fully described by the GENUS [III.33], or “number of holes,” in the surface.

Topological equivalence is rather “crude,” in the sense that it is relatively easy for two surfaces to be equivalent. As a result, the equivalence classes are parametrized by a fairly simple set: the set of all positive integers. But there are many geometrical contexts in which finer notions of equivalence are important. For example, in several contexts one wishes to regard two two-dimensional LATTICES [III.61] as equivalent if one is a rotation and enlargement of the other. Equivalence relations such as this one often lead to parameter sets that themselves have an interesting geometrical structure. Such sets are called *moduli spaces*. For details, see [IV.8] and also [V.26].

III.63 The Monster Group

THE CLASSIFICATION OF FINITE SIMPLE GROUPS [V.8] is one of the landmarks of twentieth-century mathematics. As its name suggests, it gives a complete description of all finite simple groups, which can be thought of as the building blocks for all finite groups. It states that each finite simple group belongs to one of eighteen infinite families, or else is one of twenty-six “sporadic” examples. The Monster group is the largest of the sporadic simple groups, and has 808 017 424 794 512 875 886 459 904 961 710 757 005 754 368 000 000 000 elements.

As well as having a starring role in the classification theorem, the Monster group has remarkable and deep connections with other areas of mathematics. Most notably, the smallest dimension of a faithful REPRESENTATION [IV.12] of the Monster group is 196 883, while the coefficient of $e^{2\pi iz}$ in the important and famous “ j -function” (see ALGEBRAIC NUMBERS [IV.3 §8]) is 196 884. Far from being an amusing coincidence, the fact that these two numbers differ by just 1 is a manifestation of a very deep connection between the two. See VERTEX OPERATOR ALGEBRAS [IV.13 §4.2] for further details.

The Navier–Stokes Equation

See THE EULER AND NAVIER–STOKES EQUATIONS [III.23]

III.64 Normed Spaces and Banach Spaces

It is often useful to approximate a function f by a polynomial P . For example, if you are designing a pocket calculator and want it to calculate LOGARITHMS [III.25 §4], you cannot expect it to do so exactly, since a calculator cannot handle infinitely many digits, so instead you will get it to calculate a different function $P(x)$ that approximates $\log(x)$ well. Polynomials are a good choice, because they can be built up from the basic operations of addition and multiplication. This idea raises two questions: which functions can you hope to approximate, and what counts as a good approximation?

Clearly, the answer to the second question determines the answer to the first, but there is no single right answer to the second: it is up to you what you would like to declare to be a good approximation. However, not all decisions are equally natural. Suppose that P and Q are polynomials, f and g are more general functions, and x is a real number. If $P(x)$ is close to $f(x)$ and $Q(x)$ is close to $g(x)$, then $P(x) + Q(x)$ will be close to $f(x) + g(x)$. Also, if λ is a real number and $P(x)$ is close enough to $f(x)$, then $\lambda P(x)$ will be close to $\lambda f(x)$. This informal argument suggests that the functions that we can approximate well will form a VECTOR SPACE [I.3 §2.3].

We have arrived, by one of many possible routes, at the following general situation: we are given a vector space V (consisting, in our case, of certain functions) and we would like to be able to say, in a precise way, what it is for two elements of the vector space to be *close*.

The notion of closeness is captured by METRIC SPACES [III.58], so the obvious approach is to define a metric d on the space V . Now a general principle, when putting two structures together (in this case, the linear structure of the vector space and the distance structure of the metric), is that the two structures should *relate* to one another in a natural way. In our case, there are two natural properties that one should ask for. The first is *translation invariance*. If u and v are two vectors and we translate them by adding w to both, then their distance should not change: that is, $d(u + w, v + w) = d(u, v)$. The second is that the metric should *scale correctly*. For example, if one doubles two vectors u and v , then the distance between them should double. More generally, if one multiplies u and

T&T note: difficult to fix this paragraph but will do it before CRC.

v by a scalar λ , then the distance between them should multiply by $|\lambda|$: that is, $d(\lambda u, \lambda v) = |\lambda|d(u, v)$.

If a metric has the first of these properties, then, setting $w = -u$, we find that $d(u, v) = d(0, v - u)$. It follows that if we know distances from 0, then we know all distances. Let us write $\|v\|$ instead of $d(0, v)$. Then what we have just shown is that $d(u, v) = \|v - u\|$. The expression $\|\cdot\|$ is called a *norm*, and $\|v\|$ is the *norm of v* . The following two properties of norms are easy to deduce from the fact that d is a metric that scales properly.

- (i) For any vector v , $\|v\| \geq 0$. Moreover, $\|v\| = 0$ only if $v = 0$.
- (ii) For any vector v and any scalar λ , $\|\lambda v\| = |\lambda|\|v\|$.

We also have the so-called *triangle inequality*.

- (iii) $\|u + v\| \leq \|u\| + \|v\|$ for any two vectors u and v .

This follows from translation invariance and the triangle inequality for metric spaces, since

$$\begin{aligned}\|u + v\| &= d(0, u + v) \leq d(0, u) + d(u, u + v) \\ &= d(0, u) + d(0, v) = \|u\| + \|v\|.\end{aligned}$$

In general, any function $\|\cdot\|$ on a vector space V that has properties (i)–(iii) is called a *norm on V* . A vector space with a norm on it is called a *normed space*. Given a normed space V , we can say that two vectors u and v are close if their distance $\|v - u\|$ is small.

There are many important examples of normed spaces, several of which are discussed elsewhere in this volume. One class of examples that stands out is that of *HILBERT SPACES* [III.37], which can be thought of as norms given by distances that stay the same not just when you translate but also when you rotate. Other examples are discussed in *FUNCTION SPACES* [III.29].

Let us return to the problem of how to discuss approximation by polynomials. The most commonly given answers to the two questions that arose earlier are as follows. The functions that one can approximate well are all continuous functions defined on some closed interval $[a, b]$ of real numbers. These functions form a vector space which is denoted $C[a, b]$. To make the notion of good approximation precise, we introduce a norm on this space: $\|f\|$ is defined to be the largest value of $|f(x)|$ for any x in the interval (that is, for any x between a and b). With this definition, the distance $\|f - g\|$ between two functions f and g will be small if and only if $|f(x) - g(x)|$ is small for every x in the interval. In this situation one says that f *uniformly*

approximates g . It is not obvious that every continuous function on $[a, b]$ can be uniformly approximated by a polynomial: the statement that it can is called the *Weierstrass approximation theorem*.

Here is a different way in which normed spaces arise. For most *PARTIAL DIFFERENTIAL EQUATIONS* [I.3 §5.4] it is not possible to write down a tidy formula that solves them. However, there are many techniques for proving that solutions *exist*, and they usually involve limiting arguments. For example, sometimes one can generate a sequence of functions f_1, f_2, \dots and show that these functions “converge” to some “limiting function” f , which, owing to the way we constructed the sequence f_1, f_2, \dots , must be a solution to the equation. Again, if we want to make sense of this, we must know what it is for two functions to be close, which means that the functions f_n should belong to a normed space.

How can we show that these functions converge to a limit f if we cannot already describe f ? The answer is that most interesting normed spaces, including Hilbert spaces and most important function spaces, have an additional property, called *completeness*, which guarantees, under certain conditions, that limits do indeed exist. Informally, it says that if the vectors in a sequence v_1, v_2, \dots all get very close to each other when you go far enough along the sequence, then they must converge to a limit, v , that belongs to the normed space as well. A complete normed space is known as a *Banach space*, after the Polish mathematician STEFAN BANACH [VI.84], who developed much of the general theory of such spaces. Banach spaces have many useful properties that normed spaces do not have in general: the completeness property can be thought of as ruling out pathological examples.

The theory of Banach spaces is sometimes known as *linear analysis*, since by mixing vector spaces and metric spaces it mixes linear algebra and analysis. Banach spaces arise throughout modern analysis: see, for example, the articles in this volume on *PARTIAL DIFFERENTIAL EQUATIONS* [IV.16], *HARMONIC ANALYSIS* [IV.18], and *OPERATOR ALGEBRAS* [IV.19].

III.65 Number Fields

Ben Green

A *number field K* is a “finite-degree field extension” of \mathbb{Q} , the field of rational numbers. This means that K is a *FIELD* [I.3 §2.2] that is finite dimensional when one regards it as a *VECTOR SPACE* [I.3 §2.3] over \mathbb{Q} . The following alternative description is somewhat more con-

crete. Take finitely many algebraic numbers $\alpha_1, \dots, \alpha_k$ (that is, roots of polynomials with integer coefficients) and consider the field K of all rational functions in the α_i . (In other words, K consists of numbers like $\alpha_1^2 \alpha_3 / (\alpha_2^2 + 7)$.) Then it turns out that K is a number field (the one thing that is not completely obvious is that it has finite degree over \mathbb{Q}), which we denote by $\mathbb{Q}(\alpha_1, \dots, \alpha_k)$. Conversely, every number field is of this form.

The simplest number fields are perhaps the *quadratic fields*. These are fields of the form $\mathbb{Q}(\sqrt{d}) = \{a + b\sqrt{d} : a, b \in \mathbb{Q}\}$, where d is an integer (which, it is important to stress, may be negative) that is square-free. This last condition tells us that d has no nontrivial square factors. It is there for convenience so that all the $\mathbb{Q}(\sqrt{d})$ will be distinct. (For example, $\mathbb{Q}(\sqrt{12})$, if we were to allow it, would equal $\mathbb{Q}(\sqrt{3})$, since $\sqrt{12} = 2\sqrt{3}$.) Among the other important number fields are the *cyclotomic fields*. Here we take a primitive m th root of unity ζ_m (which, for concreteness, one could take to be $e^{2\pi i/m}$) and “adjoin” it to \mathbb{Q} , obtaining the field $\mathbb{Q}(\zeta_m)$.

Why consider number fields? Historically, an important reason is that they allow us to factorize certain Diophantine equations. For example, the Ramanujan–Nagell equation $x^2 = 2^n - 7$ may be factorized as

$$(x + \sqrt{-7})(x - \sqrt{-7}) = 2^n$$

if we allow coefficients in the field $\mathbb{Q}(\sqrt{-7})$, while the Fermat equation $x^n + y^n = z^n$ is equivalent to

$$x^n = (z - y)(z - \zeta_n y) \cdots (z - \zeta_n^{n-1} y) \quad (1)$$

if we allow coefficients in the field $\mathbb{Q}(\zeta_n)$.

Before one can start thinking about whether such factorizations are useful, it is necessary to understand the notion of *integer* in a number field K . A number $\alpha \in K$ is an (algebraic) integer if it is a root of a *monic* polynomial with coefficients in \mathbb{Z} : that is, a polynomial with leading coefficient 1. For simple fields like $\mathbb{Q}(\sqrt{d})$ with d squarefree, the integers can be described quite explicitly. They are all the numbers of the form $a + b\sqrt{d}$ for integers a and b , unless $d \equiv 1 \pmod{4}$, in which case we must include more numbers: we get all numbers of the form $a + b(\frac{1}{2}(1 + \sqrt{d}))$, again for integers a and b . The set of integers in K is often denoted by \mathcal{O}_K , and it forms a RING [III.83 §1].

Unfortunately, factorizations such as (1) are not as helpful as they seem at first sight: \mathcal{O}_K turns out not to be OK, at least if one expects familiar properties of the ring \mathbb{Z} to carry over unchanged. In particular, unique factorization into primes fails to hold: for example, $2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5})$ in the field $\mathbb{Q}(\sqrt{-5})$. The

numbers on both sides are integers in this field, and it is not possible to decompose any of them any further.

Amazingly, unique factorization may be restored by embedding \mathcal{O}_K into a larger set, which consists of objects called IDEALS [III.83 §2]. There is a natural EQUIVALENCE RELATION [I.2 §2.3] that one can place on these ideals, and the number of equivalence classes, called the *class number* and written $h(K)$, is one of the most important invariants in number theory: in a certain sense, it measures “the extent to which unique factorization fails” in the number field K . (See ALGEBRAIC NUMBERS [IV.3 §7] for more details.) The fact that it is finite is one of the two basic *finiteness theorems* in algebraic number theory.

When $h(K) = 1$, the integers \mathcal{O}_K themselves enjoy unique factorization, without the need for extra ideals. This does not happen particularly often; among the fields $\mathbb{Q}(\sqrt{-d})$ with d positive and squarefree, only nine have this property, namely $d = 1, 2, 3, 7, 11, 19, 43, 67$, and 163 . The problem of determining these numbers was posed by GAUSS [VI.26] and finally solved by Heegner in 1952.

The fact that $h(\mathbb{Q}(\sqrt{-163})) = 1$ is closely related to some remarkable facts. For example, the polynomial $x^2 + x + 41$ takes prime values when $x = 0, 1, \dots, 39$ (observe that $4 \times 41 = 163 + 1$), and the number $e^{\pi\sqrt{163}}$ is within 10^{-12} of an integer.

It is a well-known open problem to decide whether or not there are infinitely many fields $\mathbb{Q}(\sqrt{d})$, $d > 0$, with class number 1. Gauss and many subsequent authors have conjectured that there are.

The second basic finiteness result in algebraic number theory is *Dirichlet’s unit theorem*. A *unit* is simply some $x \in \mathcal{O}_K$ such that there exists $y \in \mathcal{O}_K$ with $xy = 1$. The numbers 1 and -1 are always units, but there can certainly be others: for example, $17 - 12\sqrt{2}$ is a unit in $\mathbb{Q}(\sqrt{2})$ (since its reciprocal is $17 + 12\sqrt{2}$). The units form an Abelian group \mathcal{U}_K under multiplication. Dirichlet’s theorem states that this group has finite rank, which means that it is generated by finitely many of its elements.

If $d > 0$ is squarefree and if $K = \mathbb{Q}(\sqrt{d})$, then \mathcal{U}_K has rank 1. When $d \not\equiv 1 \pmod{4}$, the fact that it has rank *at least* 1 is equivalent to the statement that the Pell equation $x^2 - dy^2 = 1$ always has a nontrivial solution. This is because the Pell equation factors as $(x - y\sqrt{d})(x + y\sqrt{d}) = 1$. The unit $17 - 12\sqrt{2}$ in $\mathbb{Q}(\sqrt{2})$ corresponds to the solution $x = 17$, $y = 12$ of the equation $x^2 - 2y^2 = 1$.

For more about some of the topics discussed in this article, see FERMAT'S LAST THEOREM [V.12].

III.66 Optimization and Lagrange Multipliers

Keith Ball

1 Optimization

Soon after being introduced to calculus, most students learn of its application to *optimization*: that is, to the problem of finding the largest or smallest value of a given differentiable function, which is usually referred to as the *objective function*. A very helpful observation is that if f is an objective function that is maximized or minimized at x , then the tangent to the graph at the point $(x, f(x))$ will be horizontal, since otherwise we can find some value x' close to x for which $f(x')$ is higher. This means that we can narrow down the search for the maximum and minimum values of f by looking just at the values of $f(x)$ for which $f'(x) = 0$.

Now suppose that we have an objective function of more than one variable, such as, for example, the function

$$F(x, y) = 2x + 10y - x^2 + 2xy - 3y^2.$$

The “graph” of F is obtained by plotting the values $F(x, y)$ of F as heights above the corresponding points (x, y) of the plane, so now it is a surface instead of a curve. A smooth surface possesses not a tangent *line* at each point, but a tangent *plane*. If F has a maximum value, it will occur at a point where the tangent plane is horizontal.

The tangent plane at each point (x, y) is the graph of the linear function that best approximates F near (x, y) . For small values of h and k , $F(x + h, y + k)$ will be approximately equal to $F(x, y)$ plus a function of the form

$$(h, k) \mapsto ah + bk,$$

that is, $F(x, y)$ plus a linear function of h and k . As explained in SOME FUNDAMENTAL MATHEMATICAL DEFINITIONS [I.3 §5.3], the derivative of F at (x, y) is this linear map. The map can be represented by the pair of numbers (a, b) , which can in turn be thought of as a vector in \mathbb{R}^2 . This derivative vector is usually called the *gradient* of the function F at the point (x, y) and is written $\nabla F(x, y)$. In vector notation (writing \mathbf{x} for (x, y) and \mathbf{h} for (h, k)), the approximation to F near (x, y) is

$$F(\mathbf{x} + \mathbf{h}) \approx F(\mathbf{x}) + \mathbf{h} \cdot \nabla F. \quad (1)$$

Thus, ∇F points in the direction in which F increases most rapidly if you start at \mathbf{x} , and the magnitude of ∇F is the slope of the “graph” of F in this direction.

The components a and b of the gradient can be calculated using partial differentiation. The number a tells us how quickly $F(x, y)$ changes as we vary x while keeping y fixed: so to find a , we differentiate $F(x, y) = 2x + 10y - x^2 + 2xy - 3y^2$ with respect to x , treating y as a constant. In this case we get the partial derivative

$$a = \frac{\partial F(x, y)}{\partial x} = 2 - 2x + 2y.$$

Similarly,

$$b = \frac{\partial F(x, y)}{\partial y} = 10 + 2x - 6y.$$

Now, if we want to locate points where the tangent plane is horizontal, then we want to find the points at which the gradient is zero: that is, the points at which the vector (a, b) is the zero vector. So we solve the pair of simultaneous equations

$$2 - 2x + 2y = 0,$$

$$10 + 2x - 6y = 0$$

to get $x = 4$, $y = 3$. Thus the only candidate for the maximum is the point $(4, 3)$, where F takes the value 19. It can be checked that 19 is indeed the maximum value of F .

2 The Gradient and Contours

One of the most common ways of representing surfaces (landscapes on maps, for example) is by means of *contour lines*, or curves of constant height. In the xy -plane, we plot several curves of the form $F(x, y) = V$, for various “representative” values of V . For the function we considered earlier,

$$F(x, y) = 2x + 10y - x^2 + 2xy - 3y^2,$$

the values 0, 8, 14, 18, 19 yield the contour plot shown in figure 1. The 14 contour, for example, contains all the points at which the surface has height 14. The figure indicates that this particular surface is an elliptical hump whose peak occurs at $(4, 3)$ and has height 19.

There is a simple geometrical relationship between the contour lines and the gradient vector. The vector equation (1) shows that the direction \mathbf{h} in which F is instantaneously constant is the direction which makes the scalar product $\mathbf{h} \cdot \nabla F$ equal to 0: the direction perpendicular to ∇F . At each point, the gradient vector is perpendicular to the contour through that point. This fact underlies the method of Lagrange multipliers that we shall discuss in the next section.

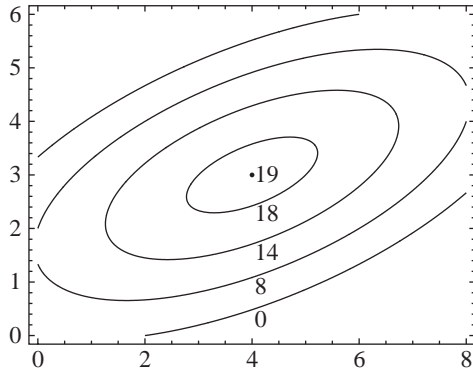


Figure 1 A contour plot.

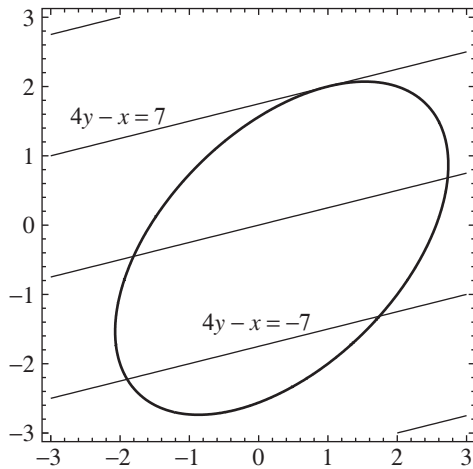


Figure 2 Constrained optimization.

3 Constrained Optimization and Lagrange Multipliers

It often happens that we are interested in the maximum or minimum value of an objective function that depends upon several variables whose values are constrained to satisfy certain equations or inequalities. Consider, for example, the following problem.

Find the maximum value of

$$F(x, y) = 4y - x$$

over all pairs (x, y) satisfying the constraint

$$G(x, y) = x^2 - xy + y^2 - x + y - 4 = 0. \quad (2)$$

Figure 2 shows the curve in the xy -plane defined by $G(x, y) = 0$ (an ellipse), and also a number of contour lines of the function $4y - x$. Our aim is to find the

largest that $4y - x$ can be if (x, y) is a point on the curve. So we want to find the largest value of V for which the corresponding contour $4y - x = V$ contains a point on the curve. The value of V increases as the lines move up the diagram, and the uppermost line that touches the curve is the one labeled $4y - x = 7$. So the maximum value we are looking for is 7, and it occurs at the point where the line $4y - x = 7$ touches the curve. It is easy to check that this point is $(1, 2)$.

How could we locate this point algebraically, rather than by drawing? The important thing to notice is that the optimizing line is tangent to the curve: the line and the curve are parallel at their common point. The line was chosen to be a contour of the function F . The curve is also a contour: the 0 contour of G . From the discussion in the previous section we know that these contours are perpendicular to the gradients of F and G , respectively (at the point in question). So the two gradient vectors are parallel to one another and are therefore multiples of one another: $\nabla F = \lambda \nabla G$, say.

We thus have a way to hunt for solutions to the constrained optimization problem

$$\text{maximize } F(x, y) \quad \text{subject to } G(x, y) = 0.$$

We look for a point (x, y) and a number λ such that

$$\nabla F(x, y) = \lambda \nabla G(x, y) \quad \text{and} \quad G(x, y) = 0. \quad (3)$$

For our example (2), the gradient equation gives two partial derivative equations,

$$-1 = \lambda(2x - y - 1), \quad 4 = \lambda(-x + 2y + 1),$$

from which we conclude that

$$x = \frac{2 + \lambda}{3\lambda}, \quad y = \frac{7 - \lambda}{3\lambda}. \quad (4)$$

If we substitute these values into the equation $G(x, y) = 0$, then we obtain

$$\frac{13(1 - \lambda^2)}{3\lambda^2} = 0,$$

which has two solutions: $\lambda = 1$ and $\lambda = -1$. If we substitute $\lambda = 1$ into (4), we get the point $(1, 2)$ where F is at its maximum. ($\lambda = -1$ yields the minimum.)

The number λ that we introduced to solve the problem is called a *Lagrange multiplier*. It is possible to reformulate the problem by defining the *Lagrangian*

$$\mathcal{L}(x, y, \lambda) = F(x, y) - \lambda G(x, y)$$

and then condensing the equations (3) into a single equation

$$\nabla \mathcal{L} = 0.$$

The reason this works is that if we differentiate \mathcal{L} with respect to λ , then we obtain $G(x, y)$, so requiring this

partial derivative to be zero is equivalent to requiring $G(x, y)$ to be zero. And asking for the other two partial derivatives to be zero is equivalent to requiring that $\nabla F = \lambda \nabla G$. The remarkable fact about this reformulation is that it has turned a *constrained* optimization problem involving x and y into an *unconstrained* problem involving x , y , and λ .

4 The General Method of Lagrange Multipliers

In real problems we may wish to optimize a function F of many variables x_1, \dots, x_n under many constraints $G_1(x_1, \dots, x_n) = 0$, $G_2(x_1, \dots, x_n) = 0$, ..., $G_m(x_1, \dots, x_n) = 0$. In this case we introduce a Lagrange multiplier for each constraint and define the Lagrangian \mathcal{L} by the formula

$$\begin{aligned} \mathcal{L}(x_1, \dots, x_n, \lambda_1, \dots, \lambda_m) \\ = F(x_1, \dots, x_n) - \sum_1^m \lambda_i G_i(x_1, \dots, x_n). \end{aligned}$$

The partial derivative of \mathcal{L} with respect to λ_i is zero if and only if $G_i(x_1, \dots, x_n) = 0$. And the partial derivatives with respect to the x_i will all be zero if and only if $\nabla F = \sum_1^m \lambda_i \nabla G_i$. This tells us that any direction that is perpendicular to all the gradients ∇G_i (and therefore lies in all their “contour hypersurfaces”) will be perpendicular to the gradient ∇F as well, so we cannot find a direction in which F increases while all the constraints are satisfied.

Problems of this kind occur frequently in economics, where the objective function F is a cost (which we are probably trying to minimize), and the constraints force us to allocate spending among different items so as to satisfy certain overall demands. For instance, we might want to minimize the total cost of supplies of various different foodstuffs that between them had to satisfy various nutritional demands. In this case, the Lagrange multipliers have an interpretation as “notional prices.” As we have just seen, at the optimum point we have an equation of the form $\nabla F = \sum_1^m \lambda_i \nabla G_i$. This tells us how much F will vary as we vary the G_i by small amounts: that is, it tells us the costs associated with increasing the various demands.

For a further use of Lagrange multipliers, see THE MATHEMATICS OF TRAFFIC IN NETWORKS [VII.4].

III.67 Orbifolds

If you take a QUOTIENT [I.3 §3.3] of the plane \mathbb{R}^2 by a group of symmetries, then you may obtain a MANIFOLD

[I.3 §6.9]. For instance, if the group consists of all translations by an integer vector, then two points (x, y) and (z, w) are equivalent if and only if $z - x$ and $w - y$ are both integers, and the quotient space is a torus. However, if you take instead the group of all rotations about the origin through a multiple of $\pi/3$, then every point apart from the origin is equivalent to exactly five others, while the origin is equivalent only to itself. The result in this case is not a manifold, because the exceptional behavior at the origin results in a singularity. However, it is a well-understood kind of singularity. An *orbifold* is, roughly speaking, just like a manifold, except that whereas manifolds are locally like \mathbb{R}^n , orbifolds are locally like quotients of \mathbb{R}^n by groups of symmetries, and can therefore have a few singularities. See ALGEBRAIC GEOMETRY [IV.7 §7] and also MIRROR SYMMETRY [IV.14 §7].

III.68 Ordinals

Loosely speaking, the ordinals are what we get if, starting with 0, we use the following two procedures. We can add 1 to whatever we have, and we can “collect together” (or “take the limit of”) whatever we have so far. So from 0 we would get 1, then 2, then 3, and so on. After all of those, we could take their “limit” (i.e., the limit of $0, 1, 2, 3, \dots$), which is called ω . Then we can add 1, obtaining $\omega + 1$, then $\omega + 2$, and so on. And then we can take the limit of all of *those*, to obtain an ordinal we could write as $\omega + \omega$. And so on. Note that this final “and so on” carries quite a lot inside it. For example, the ordinals do not just consist of finite sums of ω s and natural numbers, since we can take the limit of $\omega, \omega + \omega, \omega + \omega + \omega, \dots$, which we might call ω^2 .

Ordinals arise in two ways (which turn out to be closely related). First, they give a measure of the “size” of a *well-ordering*. A well-ordering on a set is an ordering in which every (nonempty) subset has a least element. For example, the set $\{\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \dots\} \cup \{\frac{3}{2}, \frac{5}{3}, \frac{7}{4}, \dots\}$ in the reals is well-ordered, while the set $\{\dots, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}\}$ is not. The first set is *order isomorphic* to the ordinals less than $\omega + \omega$, meaning that there is a bijection that preserves the order. So one says that that set has *order type* $\omega + \omega$.

Ordinals also commonly arise when one wishes to index transfinite processes. Here “transfinite” means “going beyond finite.” As an example, suppose that we wish to “count, in increasing order” the elements of the well-ordered set above. How would we do it? We would start with $\frac{1}{2}$, then $\frac{2}{3}$, then $\frac{3}{4}$, and so on. But, at the end of

all time, we would still not have reached elements like $\frac{3}{2}$ or $\frac{5}{3}$. So we would start again: “at time ω ” we would count $\frac{3}{2}$, then at time $\omega + 1$ we would count $\frac{5}{3}$, and so on. Thus our counting is complete by time $\omega + \omega$.

For a more detailed explanation of ordinals, including more examples and more on how they arise in mathematics, see SET THEORY [IV.1 §2].

III.69 The Peano Axioms

Everyone knows what the natural numbers are: 0, 1, 2, 3, and so on. But how would we make that “and so on” precise? Can we look at the way that we reason about natural numbers and isolate a few basic principles, or *axioms*, whose consequences do complete justice to our intuitive picture of what the natural numbers should be? To put it another way, when we are proving something about the natural numbers, what assumptions do we need in order to get started?

To answer this question, let us strip things down to the bare minimum: we have an object called 0, and an operation s , called the *successor function*, which we think of intuitively as “adding 1.” In this pared-down language, we would like to say two things: that all the numbers $0, s(0), s(s(0)), \dots$ are distinct natural numbers, and that there are no others.

One simple way is to use the following two axioms. The first says that 0 is not a successor:

- (i) For all x , $s(x) \neq 0$.

The second says that distinct elements stay distinct when you take their successors:

- (ii) For all x and y , if $x \neq y$, then $s(x) \neq s(y)$.

Note that this implies, for example, that $s(s(s(0))) \neq s(0)$, for if they *were* equal, then, from rule (ii), we could deduce that $s(s(0)) = 0$, contradicting rule (i).

Now, how can we say that there are no other natural numbers? One would like to say that, for every x , either $x = 0$ or $x = s(0)$ or $x = s(s(0))$ or \dots , but that is an infinitely long statement, and those are definitely not allowed. After the failure of that very natural attempt, one might guess that there is no way to achieve the goal, but in fact there is a brilliant solution: induction. Here is an axiom that expresses the principle of induction.

- (iii) Let A be any subset of the natural numbers with the following properties: $0 \in A$, and $s(x) \in A$ whenever $x \in A$. Then A must be the set of all natural numbers.

Note that this does express our intuitive idea that there are no “extra” natural numbers, since we can take A to be the set of all the numbers $0, s(0), s(s(0)), \dots$ that were on our list.

Rules (i), (ii), and (iii) are called the *Peano axioms* for the natural numbers. As explained above, they “characterize” the natural numbers, in the sense that all reasoning about the natural numbers may be reduced or rewritten in such a way that the only assumptions one needs are the Peano axioms.

There is a related system used in logic, called the *first-order Peano axioms*. The idea here is that we want to express the Peano axioms in the language of first-order logic. This means that we are allowed variables (that are interpreted as ranging over the natural numbers), as well as the symbols 0 and s , logical connectives, and the like, but nothing more: so there is no “member of” symbol, and no sets are allowed. (However, for technical reasons one does allow symbols for “plus” and “times.”)

To give an idea of what is allowed and what is not, consider the statements “there are infinitely many perfect squares” and “every infinite set of positive integers contains either infinitely many odd numbers or infinitely many even numbers.” With a little effort, we can express the first of these statements in first-order logic, as follows:

$$(\forall m)(\exists n)(\exists x) \quad xx = m + n.$$

In words, this says that for every m you can find a perfect square of the form $m + n$ (which is how we express the fact that it is larger than m). However, in order to express the second statement, we find ourselves wanting to write $(\forall A)$, where A ranges over all possible *subsets* of the natural numbers, rather than all possible *elements*: this is the main thing that is not allowed in first-order logic.

By this criterion, rules (i) and (ii) are fine, but rule (iii) is not. Instead, we have to use an “axiom scheme,” which is an infinite set of axioms, one for each first-order statement $p(x)$. So our version of rule (iii) is this: for each statement $p(x)$, we have an axiom saying that if $p(0)$ is true, and $p(x)$ implies $p(s(x))$, then $p(x)$ is true for all x .

Note that these axioms do not have the full strength of the usual Peano axioms. For instance, there are only countably many possible formulas $p(x)$, whereas there are uncountably many sets A . It turns out that in fact there are “nonstandard” models of these axioms, meaning structures other than the natural numbers that satisfy the axioms of first-order Peano arithmetic.

Actually, one also allows *parameters* in the statements $p(x)$; for example, $p(x)$ could be the statement “there exists z with $x = y + z$,” which would correspond to the set of all natural numbers greater than or equal to y , and would therefore depend on y . And one also adds some axioms saying how plus and times behave (for example, commutativity of addition). This whole collection of axioms is known as Peano arithmetic, or PA for short.

See MODEL THEORY [IV.2] for more on some of the topics discussed in this article.

III.70 Permutation Groups

Martin W. Liebeck

Let S be a set. A *permutation* of S is a function from S to S that is both injective and surjective—in other words, a function that “rearranges” the elements of S . For example, if $S = \{1, 2, 3\}$, then the function $a : S \rightarrow S$ that sends 1 to 3, 2 to 1, and 3 to 2 is a permutation of S ; so is the function b that sends 1 to 3, 2 to 2, and 3 to 1; whereas the function c that sends 1 to 3, 2 to 1, and 3 to 1 is not a permutation. An example of a permutation of the set of real numbers \mathbb{R} is the function $x \mapsto 8 - 2x$.

From the point of view of finite group theory, the most important permutations to study are those of the set $I_n = \{1, 2, \dots, n\}$, where n is a positive integer. Let S_n denote the set of all permutations of I_n . So, for example, the permutations a and b defined in the previous paragraph lie in S_3 . To count how many permutations there are altogether in S_n , observe that, for a permutation $f : I_n \rightarrow I_n$, there are n choices for $f(1)$, then $n - 1$ choices for $f(2)$ (we can choose anything different from $f(1)$), then $n - 2$ for $f(3)$, and so on, until we have just 1 choice for $f(n)$. Therefore the total number of permutations in S_n is $n(n - 1)(n - 2) \cdots 1 = n!$.

If f and g are permutations of a set S , their composition $f \circ g$ is defined by $f \circ g(s) = f(g(s))$ for all $s \in S$, and it is quite easy to see that $f \circ g$ is also a permutation of S . It is usual to drop the “ \circ ” symbol and write just fg instead of $f \circ g$. For example, if $a, b \in S_3$ are as in the first paragraph, then $ab \in S_3$ sends 1 to 2, 2 to 1, and 3 to 3, while ba sends 1 to 1, 2 to 3, and 3 to 2. Notice that $ab \neq ba$.

For any set S , the *identity* function $\iota : S \rightarrow S$, defined by $\iota(s) = s$ for all $s \in S$, is a permutation of S ; and if f is a permutation of S , then there is an *inverse* permutation f^{-1} that sends everything back to where it came from and therefore satisfies $ff^{-1} = f^{-1}f = \iota$. For example, the inverse of the above permutation $a \in S_3$

is the permutation that sends 1 to 2, 2 to 3, and 3 to 1. Also, for any permutations f, g, h of S , we have $f(gh) = (fg)h$, since both sides send any $s \in S$ to $f(g(h(s)))$.

Thus, the set of all permutations of S , together with the BINARY OPERATION [I.2 §2.4] of composition, satisfies the axioms for a GROUP [I.3 §2.1]. In particular, S_n is a finite group of size $n!$, known as the *symmetric group of degree n* .

There is a neat way of representing permutations succinctly, known as the *cycle notation*. It is best explained with an example. Let $d \in S_6$ be the permutation $1 \mapsto 3, 2 \mapsto 5, 3 \mapsto 6, 4 \mapsto 4, 5 \mapsto 2, 6 \mapsto 1$. We can represent this more economically by writing $1 \mapsto 3 \mapsto 6 \mapsto 1$, and $4 \mapsto 4$. We say the symbols 1, 3, 6 form a *cycle* of d (of length 3); similarly, 2, 5 form a cycle of length 2, and 4 a cycle of length 1. We then compress our notation even further and write $d = (1\ 3\ 6)(2\ 5)(4)$, indicating that each number 1, 3, 6 in the first cycle is sent to the next one, except for the last which is sent back to the first, and likewise for the second and third cycles. This is the cycle notation for d ; notice that the cycles have no symbols in common—they are called *disjoint* cycles. It is not too hard to see that every permutation in S_n can be expressed as a product of disjoint cycles; this is what we mean by the cycle notation for a permutation. For example, in cycle notation, the six permutations of S_3 are $\iota, (1\ 2)(3), (1\ 3)(2), (2\ 3)(1), (1\ 2\ 3)$, and $(1\ 3\ 2)$. (The permutations a and b in the first paragraph are $(1\ 3\ 2)$ and $(1\ 3)(2)$, respectively.) You might like to while away a few minutes by working out the multiplication table of S_3 .

The *cycle-shape* of a permutation g is the sequence of numbers we get by writing down the lengths of the disjoint cycles in the cycle notation for g , in decreasing order. For example, the cycle-shape of the permutation $(1\ 6\ 3)(2\ 4)(5\ 8)(7)(9)$ in S_9 is $(3, 2, 2, 1, 1)$, or more succinctly $(3, 2^2, 1^2)$.

One can define the *powers* of a permutation $f \in S_n$ in a natural way—namely, $f^1 = f, f^2 = ff, f^3 = f^2f$, and so on. For example, if $e = (1\ 2\ 3\ 4) \in S_4$, then $e^2 = (1\ 3)(2\ 4)$, $e^3 = (1\ 4\ 3\ 2)$, and $e^4 = \iota$. The *order* of a permutation $f \in S_n$ is defined to be the smallest positive integer r such that $f^r = \iota$: that is, the smallest number of times we have to do f to send everything back to where it came from. So the order of the 4-cycle e above is 4. In general, the order of an r -cycle (i.e., a cycle of length r) is equal to r , and the order of a permutation in cycle notation is equal to the least common multiple of the lengths of the (disjoint) cycles.

It is often useful to be able to work out the order of a permutation. Here is one such instance. Suppose we shuffle a pack of eight cards in the following way: the pack is divided into two equal parts and then “interlaced,” so that if the original order was $1, 2, 3, 4, \dots$, the new order is $1, 5, 2, 6, \dots$. How many times must this shuffle be repeated before the cards are again in the original order? Well, the shuffle gives the permutation of the eight card positions sending 1 to 1, 2 to 5, 3 to 2, 4 to 6, and so on, which in cycle notation is $(1)(2\ 5\ 3)(4\ 6\ 7)(8)$. This has order 3, so the cards return to their original order after three shuffles. Things get quite interesting if we consider the same problem for different numbers of cards—you might like to try it yourself with fifty-two cards, for instance.

There is one slightly more subtle aspect of permutations which is important for group theory: namely, the theory of *even* and *odd* permutations. Again, this is best illustrated by example. Take $n = 3$, and let x_1, x_2, x_3 be three variables. Let us think of the permutations in S_3 as moving these variables around rather than the numbers 1, 2, and 3. So, for instance, we shall take the permutation $(1\ 3\ 2)$ to send x_1 to x_3 , x_2 to x_1 , and x_3 to x_2 . Now let Δ be the expression $\Delta = (x_1 - x_2)(x_1 - x_3)(x_2 - x_3)$. We can apply permutations in S_3 to Δ in an obvious way: for example, $(1\ 2\ 3)$ sends Δ to $(x_2 - x_3)(x_2 - x_1)(x_3 - x_1)$. Notice that this is just the expression for Δ with two of the brackets, $(x_1 - x_2)$ and $(x_1 - x_3)$, reversed. So $(1\ 2\ 3)$ sends Δ to Δ . However, if we apply $(1\ 2)(3)$ to Δ , we get $(x_2 - x_1)(x_2 - x_3)(x_1 - x_3) = -\Delta$. You can see that each permutation in S_3 sends Δ to either $+\Delta$ or $-\Delta$. Call those permutations that send Δ to $+\Delta$ *even permutations* and those that send Δ to $-\Delta$ *odd permutations*. Check that ι , $(1\ 2\ 3)$, and $(1\ 3\ 2)$ are even, while $(1\ 2)(3)$, $(1\ 3)(2)$, and $(2\ 3)(1)$ are odd.

The definition of even and odd permutations for general n is very similar to this example. Let x_1, \dots, x_n be variables, and take the permutations in S_n to move these variables around rather than the symbols $1, 2, \dots, n$. Define Δ to be the product of all $x_i - x_j$ for $i < j$. Just as in the example, we can apply any permutation $g \in S_n$ to Δ , and the result will be either $+\Delta$ or $-\Delta$. Define the *signature* of g to be the number $\text{sgn}(g) \in \{+1, -1\}$ such that $g(\Delta) = \text{sgn}(g)\Delta$. This defines the signature function $\text{sgn} : S_n \rightarrow \{+1, -1\}$. Then a permutation $g \in S_n$ is *even* if $\text{sgn}(g) = +1$, and is *odd* if $\text{sgn}(g) = -1$.

It follows easily from the definition that

$$\text{sgn}(gh) = \text{sgn}(g) \text{sgn}(h)$$

for any $g, h \in S_n$, and also that the signature of any 2-cycle is -1 . Since an r -cycle $(a_1\ a_2\ \dots\ a_r)$ can be expressed as a product $(a_1\ a_r)(a_1\ a_{r-1}) \dots (a_1\ a_2)$ of 2-cycles, the signature of the r -cycle is $(-1)^{r-1}$. Hence, if $g \in S_n$ has cycle-shape (r_1, r_2, \dots, r_k) , then

$$\text{sgn}(g) = (-1)^{r_1-1}(-1)^{r_2-1} \dots (-1)^{r_k-1}.$$

This makes it easy to work out the signature of any permutation. For example, the even permutations in S_5 are those that have cycle-shape (1^5) , $(2^2, 1)$, $(3, 1^2)$, or (5) . If you count these, you will find that there are sixty even permutations in S_5 altogether, which is exactly half of the total of $5! = 120$ permutations in S_5 . In general, the number of even permutations in S_n is $\frac{1}{2}n!$.

So what is the point of this complicated definition? The answer is that the set of all even permutations in S_n forms a subgroup of size $\frac{1}{2}n!$, known as the *alternating group of degree n* , and written as A_n . The alternating groups are very important examples of finite groups, because of the fact that, for $n \geq 5$, A_n is a *simple* group—that is, its only NORMAL SUBGROUPS [I.3 §3.3] are the identity subgroup and A_n itself (see THE CLASSIFICATION OF FINITE SIMPLE GROUPS [V.8]). For example, A_5 is a simple group of size 60, and in fact is the smallest non-Abelian finite simple group.

III.71 Phase Transitions

If you heat up a block of ice, then it turns into water. This very familiar phenomenon is actually rather mysterious, because it shows that the properties of the chemical H_2O do not depend continuously on temperature: the block of ice goes straight from a solid to a liquid, rather than doing so by a process of gradual softening.

This is an example of a *phase transition*. Phase transitions tend to occur in systems that involve a large number of particles with “local” interactions—that is, where the behavior of one particle is directly influenced only by the particles in its immediate vicinity.

Such systems can be modeled mathematically, and the study of these models belongs to the area known as *statistical physics*. For further discussion of such models, see PROBABILISTIC MODELS OF CRITICAL PHENOMENA [IV.26].

III.72 π

What makes one number more fundamental and important, mathematically speaking, than another? Why, for

instance, would almost everybody agree that 2 is more important than $\frac{43}{32}$? One possible answer is that what really matters about a number is its properties, and in particular any interesting properties it might have that distinguish it from all other numbers. Of course, we now have to decide what counts as an interesting property: for example, why do we not regard it as interesting that $\frac{43}{32}$ is the only number that gives you $\frac{43}{16}$ when you double it? An obvious reason is that there is an analogous property for *every* number x you might care to choose: x is the only number that gives you $2x$ when you double it. By contrast, the property “is the smallest prime number” does not mention any specific number and is easily stated in terms of a concept, that of “prime number,” whose importance is itself easy to explain. This property must apply to exactly one number, so it is likely that that number will have an important part to play in mathematics, and indeed it does. (As it happens, $\frac{43}{32}$ is conjectured to be an important critical exponent in statistical physics, which means that it *can* be singled out as an interesting number, though still nothing like as fundamental as 2.)

Everybody agrees that π is one of the most important numbers in mathematics, and it is easy to justify this assessment by the criterion of the previous paragraph, because π has an abundance of properties—so many that when π appears unexpectedly in a calculation, one is not unduly surprised. For example, the following is a famous theorem of Euler:

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \frac{1}{25} + \cdots = \frac{\pi^2}{6}.$$

What on earth, one might wonder, has π to do with adding up reciprocals of squares? This is a perfectly legitimate question, but the idea that there could in principle be a connection is not, to an experienced mathematician, a surprise. A very common way to prove mathematical identities is to show that the two sides of the identity are different ways of evaluating the same quantity. In this case, one can use a basic fact from FOURIER ANALYSIS [III.27], known as *Parseval's identity*, which states the following. If $f: \mathbb{R} \rightarrow \mathbb{C}$ is a periodic function with period 2π , and for every integer n (positive or negative) we define its n th Fourier coefficient a_n by the formula

$$a_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{inx} dx,$$

then

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} |f(x)|^2 dx = \sum_{n=-\infty}^{\infty} |a_n|^2.$$

If you now take as f the function that is 1 whenever x is between $(2n - \frac{1}{2})\pi$ and $(2n + \frac{1}{2})\pi$ for some integer n , and 0 otherwise, then you find that the left-hand side works out as $\frac{1}{2}$. You also find, after a small calculation, that $|a_n|^2 = 1/\pi n^2$ when n is odd, that $|a_0|^2 = \frac{1}{4}$, and that $|a_n|^2 = 0$ whenever n is even and nonzero. Therefore,

$$\frac{1}{2} = \frac{1}{4} + \frac{1}{\pi^2} \sum_{n \text{ odd}} \frac{1}{n^2}.$$

Bearing in mind that $n^2 = (-n)^2$, we can deduce easily that

$$\frac{\pi^2}{8} = 1 + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \cdots.$$

This closely resembles the identity we were trying to prove, which we can get by noticing that the right-hand side is equal to $\sum_n 1/n^2 - \sum_n 1/(2n)^2$, which is three quarters of $\sum_n 1/n^2$. Therefore, $\sum_n 1/n^2 = \pi^2/6$.

Now we have a reason for the appearance of π : it comes up in the formula for the Fourier coefficients. What is more, its appearance there can be explained as well. A periodic function on \mathbb{R} is more naturally thought of as a function defined on the unit circle. The Fourier coefficient a_n is a certain average defined on the unit circle, so we have to divide by the length of the circle, which is 2π .

What, then, is π ? Well, we have just seen what is perhaps the most elementary definition: it is the ratio of the circumference of a circle to its diameter. But what makes π so interesting is that it has many different defining properties. Here are a few more of them.

- (i) Define a function $\sin x$ to be equal to the sum of the power series

$$x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots.$$

Then π is the smallest positive number x such that $\sin x = 0$. (For more on $\sin x$, see TRIGONOMETRIC FUNCTIONS [III.94].)

(ii) $\pi = \int_{-1}^1 \frac{dx}{\sqrt{1-x^2}}.$

(iii) $\frac{\pi}{2} = \int_{-1}^1 \sqrt{1-x^2} dx.$

(iv) $\frac{\pi}{4} = \left(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots\right).$

(v) $\sqrt{2\pi} = \int_{-\infty}^{\infty} e^{-x^2} dx.$

(vi) $\pi = \sum_{k=0}^{\infty} \frac{1}{16^k} \left(\frac{4}{8k+1} - \frac{2}{8k+4} - \frac{1}{8k+5} - \frac{1}{8k+6} \right).$

The integrals on the right-hand sides of the second and third properties above are expressions for half the circumference of the unit circle and half its area, respectively. So those definitions are analytical expressions of the geometrical facts that a unit circle has circumference 2π and area π , respectively.

The fifth property tells us what constant to put in front of e^{-x^2} to make it into the famous NORMAL DISTRIBUTION [III.73 §5]. (Why should π come into it? One can give several reasons. One is that the function e^{-x^2} has a special role in Fourier analysis, and so does π . Another fundamental property of e^{-x^2} is that the function $f(x, y) = e^{-(x^2+y^2)}$ is *rotationally invariant*, and rotations involve circles, which involve π .)

The last formula above is a remarkable recent discovery of David Bailey, Peter Borwein, and Simon Plouffe. The presence of the factor $1/16^k$ leads to a way of calculating hexadecimal digits of π (that is, digits to base 16), without needing to work out all the earlier digits first. It has been used to work out digits that are astonishingly far along the hexadecimal expansion: for example, it is known that the trillionth hexadecimal digit is 8.

A fact that seems paradoxical to many nonmathematicians is that a number as natural as π turns out to be IRRATIONAL, and also TRANSCENDENTAL [III.43]. However, this is not surprising at all: the defining properties of π are simple, but they do not lead to solutions of polynomial equations, so it would be extraordinary if π were *not* transcendental. Similarly, it would be a major surprise if one could find any pattern in the decimal digits of π . Indeed, π is conjectured to be *normal to base 10*, meaning that every sequence of digits occurs with about the frequency you would expect: for example, if you look at pairs of consecutive digits, then you expect 35 to occur about a hundredth of the time. However, this conjecture seems to be very hard, and it has not even been proved that the decimal expansion of π contains all the digits from 0 to 9 infinitely often.

III.73 Probability Distributions

James Norris

1 Discrete Distributions

When we toss a coin, we have no idea whether it will land heads or tails. However, there is a different sense in which the behavior of the coin is highly predictable: if it is tossed many times, then the proportion of heads is very likely to be close to $\frac{1}{2}$.

In order to study this phenomenon mathematically, we need to model it, and this is done by defining a *sample space*, which represents the set of possible outcomes, and a *probability distribution* on that space, which tells you their probabilities. In the case of a coin, the natural sample space is the set $\{H, T\}$, and the obvious distribution assigns the number $\frac{1}{2}$ to each element. Alternatively, since we are interested in the number of heads, we could use the set $\{0, 1\}$ instead: after one toss, there is a probability of $\frac{1}{2}$ that the number of heads is 0 and a probability of $\frac{1}{2}$ that it is 1. More generally, a (discrete) sample space is simply a set Ω , and a probability distribution on Ω is a way of assigning a nonnegative real number to each element of Ω in such a way that the sum of all these numbers is 1. The number assigned to a particular element of Ω is then interpreted as the probability that some corresponding outcome will occur, the total probability being 1.

If Ω is a set of size n , then the *uniform distribution* on Ω is the probability distribution that assigns a probability of $1/n$ to each element of Ω . However, it is often more appropriate to assign different probabilities to different outcomes. For example, given any real number p between 0 and 1, the *Bernoulli distribution with parameter p* on the set $\{0, 1\}$ is the distribution that assigns the number p to 1 and $1 - p$ to 0. This can be used to model the toss of a biased coin.

Suppose now that we toss an unbiased coin n times. If we are interested in the outcome of every toss, then we would choose the sample space consisting of all possible sequences of 0s and 1s of length n . For instance, if $n = 5$, a typical element of the sample space is 01101. (This particular element represents the outcome tails, heads, heads, tails, heads, in that order.) Since there are 2^n such sequences and they are all equally likely, the appropriate distribution on this space will be the uniform one, which assigns a probability of $1/2^n$ to each sequence.

But what if we are interested not in the particular sequence of heads and tails but just in the *total number of heads*? In that case, we could take as our sample space the set $\{0, 1, 2, \dots, n\}$. The probability that the total number of heads is k is 2^{-n} times the number of sequences of 0s and 1s that contain exactly k 1s. This number is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

so the probability we assign to k is $p_k = \binom{n}{k} 2^{-n}$.

More generally, for a sequence of n independent experiments, each with the same probability p of suc-

cess, the probability of a given sequence of k successes and $n - k$ failures is $p^k(1 - p)^{n-k}$. So, the probability of having exactly k successes is $p_k = \binom{n}{k} p^k(1 - p)^{n-k}$. This is called the *binomial distribution* with parameters n and p . It models the number of heads if you toss a biased coin n times, for example.

Suppose we perform such experiments for as long as we need to in order to obtain one success. When k experiments are performed, the probability of getting $k - 1$ failures followed by a success is $p_k = (1 - p)^{k-1} p$. Therefore, this formula gives us the distribution of the number of experiments up to the first success. It is called the *geometric distribution* of parameter p . In particular, the number of tosses of a fair coin needed to get the first head has a geometric distribution of parameter $\frac{1}{2}$. Notice that our sample space is now the set of all nonnegative integers—in particular, it is infinite. So in this case the condition that the probabilities add up to 1 is requiring that a certain infinite series (the series $\sum_{k=1}^{\infty} p_k$) converges to 1.

Now let us imagine a somewhat more complicated experiment. Suppose we have a radioactive source that occasionally emits an alpha particle. It is often reasonable to suppose that these emissions are independent and equally likely to occur at any time. If the average number of emissions per minute is λ , say, then what is the probability that during any given minute there will be k particles emitted?

One way to think about this question is to divide up the minute into n equal intervals, for some large n . If n is large enough, then the probability of two emissions occurring in the same interval is so small that it can be ignored, and therefore, since the average number of emissions per minute is λ , the probability of an emission during any given interval must be approximately λ/n . Let us call this number p . Since the emissions are independent, we can now regard the number of emissions as the number of successes when we do n trials, each with probability p of success. That is, we have the binomial distribution with parameters n and p , where $p = \lambda/n$.

Notice that as n gets larger, p gets smaller. Also, the approximations just made become better and better. It is therefore natural to let n tend to infinity and study the resulting “limiting distribution.” It can be checked that, in the limit as $n \rightarrow \infty$, the binomial probabilities converge to $p_k = e^{-\lambda} \lambda^k / k!$. These numbers define a distribution on the set of all nonnegative integers, known as the *Poisson distribution* of parameter λ .

2 Probability Spaces

Suppose that I throw a dart at a dartboard. Not being very good at darts, I am not able to say very much about where the dart will land, but I can at least try to model it probabilistically. The obvious sample space to take consists of a circular disk, the points of which represent where the dart lands. However, now there is a problem: if I look at any particular point in the disk, the probability that the dart will land at precisely that point is zero. So how do I define a probability distribution?

A clue to the answer lies in the fact that it seems to be perfectly easy to make sense of a question such as “What is the probability that I will hit the bull’s-eye?” In order to hit the bull’s-eye, the dart has to land in a certain region of the board, and the probability of this happening does not have to be zero. It might, for instance, be equal to the area of the bull’s-eye region divided by the total area of the board.

What we have just observed is that even if we cannot assign probabilities to individual *points* in the sample space, we can still hope to give probabilities to *subsets*. That is, if Ω is a sample space and A is a subset of Ω , we can try to assign a number $\mathbb{P}(A)$ between 0 and 1 to the set A . This represents the probability that the random outcome belongs to the set A , and can be thought of as something like a notion of “mass” for the set A .

For this to work, we need $\mathbb{P}(\Omega)$ to be 1 (since the probability of getting *something* in the sample space must be 1). Also, if A and B are disjoint subsets of Ω , then $\mathbb{P}(A \cup B)$ should be $\mathbb{P}(A) + \mathbb{P}(B)$. From this it follows that if A_1, \dots, A_n are all disjoint, then $\mathbb{P}(A_1 \cup \dots \cup A_n)$ is equal to $\mathbb{P}(A_1) + \dots + \mathbb{P}(A_n)$. Actually, it turns out to be important that this should be true not just for finite unions but even for COUNTABLY INFINITE [III.11] ones as well. (Related to this point is the fact that one does not attempt to define $\mathbb{P}(A)$ for *every* subset A of Ω but just for MEASURABLE SUBSETS [III.57]. For our purposes, it is sufficient to regard $\mathbb{P}(A)$ as given whenever A is a set we can actually define.)

A *probability space* is a sample space Ω together with a function \mathbb{P} , defined on all “sensible” subsets A of Ω , that satisfies the conditions mentioned in the previous two paragraphs. The function \mathbb{P} itself is known as a *probability measure* or *probability distribution*. The term *probability distribution* is often preferred when we specify \mathbb{P} concretely.

PUP: while I too find it confusing, Tim has convinced me that the use of ‘biased coin’ and ‘unbiased coin’ in this article is correct. OK?

3 Continuous Probability Distributions

There are three particularly important distributions defined on subsets of \mathbb{R} , of which two will be discussed in this section. The first is the *uniform distribution* on the interval $[0, 1]$. We would like to capture the idea that “all points in $[0, 1]$ are equally likely.” In view of the problems mentioned above, how should we do this?

A good way is to take seriously the “mass” metaphor. Although we cannot calculate the mass of an object by adding up the masses of all the infinitely small points that make up the object, we can assign to those points a *density* and integrate it. That is exactly what we shall do here. We assign a *probability density* of 1 to each point in the interval $[0, 1]$. Then we determine the probability of a subinterval, $[\frac{1}{3}, \frac{1}{2}]$ say, by calculating the integral $\mathbb{P}([\frac{1}{3}, \frac{1}{2}]) = \int_{1/3}^{1/2} 1 \, dx = \frac{1}{6}$. More generally, the probability associated with an interval $[a, b]$ will just be its length $b - a$. The probability of a union of intervals will then be the sum of the lengths of those intervals, and so on.

This “continuous” uniform distribution sometimes arises naturally from requirements of symmetry, just like its discrete counterpart. It can also arise as a limiting distribution. For instance, suppose that a hermit lives deep in a cave, away from any clocks or sources of natural light, and that each “day” he spends lasts for a random length of time between twenty-three and twenty-five hours. To start with, he will have some idea of what the time is, and be able to make statements such as, “I’m having lunch now, so it’s probably light outside,” but after a few weeks of this regime, he will no longer have any idea: any outside time will be just as likely as any other.

Now let us look at a rather more interesting density function, which depends on the choice of a positive constant λ . Consider the density function $f(x) = \lambda e^{-\lambda x}$, defined on the set of all nonnegative real numbers. To work out the probability associated with an interval $[a, b]$, we now calculate

$$\int_a^b f(x) \, dx = \int_a^b \lambda e^{-\lambda x} \, dx = e^{-\lambda a} - e^{-\lambda b}.$$

The resulting probability distribution is called the *exponential distribution with parameter λ* . The exponential distribution is appropriate if we are modeling the time T of a spontaneous event, such as the time it takes for a radioactive nucleus to decay, or for the next spam email to arrive. The reason for this is based on the assumption of *memorylessness*: for example, if we know that the nucleus remains intact at time s , the

probability that it will remain intact until a later time $s + t$ is the same as the original probability that it would remain intact to time t . Let $G(t)$ represent the probability that the nucleus remains intact up to time t . Then the probability that it remains intact up to time $s + t$ given that it has remained intact up to time s is $G(s + t)/G(s)$, so this has to equal $G(t)$. Equivalently, $G(s + t) = G(s)G(t)$. The only decreasing functions that have this property are EXPONENTIAL FUNCTIONS [III.25], that is, functions of the form $G(t) = e^{-\lambda t}$ for some positive λ . Since $1 - G(t)$ represents the probability that the nucleus decays before time t , this should equal $\int_0^t f(x) \, dx$, from which it is easy to deduce that $f(x) = \lambda e^{-\lambda x}$.

We shall come to the third, and most important, distribution below.

4 Random Variables, Mean, and Variance

Given a probability space, an *event* is defined to be a (sufficiently nice) subset of that space. For example, if the probability space is the interval $[0, 1]$ with the uniform distribution, then the interval $[\frac{1}{2}, 1]$ is an event: it represents a randomly chosen number between 0 and 1 turning out to be at least $\frac{1}{2}$. It is often useful to think not just about random events, but also about random *numbers* associated with a probability space. For example, let us look once again at a sequence of tosses of a biased coin that has probability p of coming up heads. The natural sample space associated with this experiment is the set Ω of all sequences ω of 0s and 1s. Earlier, we showed that the probability of obtaining k heads is $p_k = \binom{n}{k} p^k (1 - p)^{n-k}$, and we described that as a distribution on the sample space $\{0, 1, 2, \dots, n\}$. However, it is in many ways more natural, and often far more convenient, to regard the original set Ω as the sample space and to define a function X from Ω to \mathbb{R} to represent the number of heads: that is, $X(\omega)$ is the number of 1s in the sequence ω . We then write

$$\mathbb{P}(X = k) = p_k = \binom{n}{k} p^k (1 - p)^{n-k}.$$

A function like this is called a *random variable*. If X is a random variable and it takes values in a set Y , then the *distribution* of X is the function P defined on subsets of Y by the formula

$$P(A) = \mathbb{P}(X \in A) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\}).$$

It is not hard to see that P is indeed a probability distribution on Y .

For many purposes, it is enough to know the distribution of a random variable. However, the notion of a random variable defined on a sample space captures our intuition of a random quantity, and it allows us to ask further questions. For example, if we were to ask for the probability that there were k heads given that the first and last tosses had the same outcome, then the distribution of X would not provide the answer, whereas our richer model of regarding X as a function defined on sequences would do so. Furthermore, we can talk of *independent* random variables, X_1, \dots, X_n say, meaning that the subset of Ω where $X_i(\omega) \in A_i$ for all i has probability given by the product $\mathbb{P}(X_1 \in A_1) \times \dots \times \mathbb{P}(X_n \in A_n)$ for all possible sets of values A_i .

Associated with a random variable X are two important numbers that begin to characterize it, called the *mean* or *expectation* $\mathbb{E}(X)$ and the *variance* $\text{var}(X)$. Both these numbers are determined by the distribution of X . If X takes integer values, with distribution $\mathbb{P}(X = k) = p_k$, then

$$\mathbb{E}(X) = \sum_k k p_k, \quad \text{var}(X) = \sum_k (k - \mu)^2 p_k,$$

where $\mu = \mathbb{E}(X)$. The mean tells us how big X is on average. The variance, or more precisely its square root, the *standard deviation* $\sigma = \sqrt{\text{var}(X)}$, tells us how far away X lies, typically, from its mean. It is not hard to derive the following useful alternative formula for the variance:

$$\text{var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$

To understand the meaning of the variance, consider the following situation. Suppose that one hundred people take an exam and you are told that their average mark is 75%. This gives you some useful information, but by no means a complete picture of how the marks are distributed. For example, perhaps the exam consisted of four questions of which three were very easy and one almost impossible, so that all the marks were clustered around 75%. Or perhaps about fifty people got full marks and fifty got around half marks. To model this situation let the sample space Ω consist of the hundred people and let the probability distribution be the uniform distribution. Given a random person ω , let $X(\omega)$ be that person's mark. Then in the first situation, the variance will be small, since almost everybody's mark is close to the mean of 75%, whereas in the second it is close to $25^2 = 625$, since almost everybody's mark was about 25 away from the mean. Thus, the variance

helps us to understand the difference between the two situations.

As we discussed at the start of this article, it is known from experience that the “expected” number of heads in a sequence of n tosses of a fair coin is around $\frac{1}{2}n$, in the sense that the proportion is usually close to $\frac{1}{2}$. It is not hard to work out that, if X models the number of heads in n tosses, that is, if X is binomially distributed with parameters n and $\frac{1}{2}$, then $\mathbb{E}(X) = \frac{1}{2}n$. The variance of X is $\frac{1}{4}n$, so the natural distance scale with which to measure the spread of the distribution is $\sigma = \frac{1}{2}\sqrt{n}$. This allows us to see that X/n is close to $\frac{1}{2}$ with probability close to 1 for large n , in accordance with experience.

More generally, if X_1, X_2, \dots, X_n are independent random variables, then $\text{var}(X_1 + \dots + X_n) = \text{var}(X_1) + \dots + \text{var}(X_n)$. It follows that if all the X_i have the same distribution with mean μ and variance σ^2 , then the variance of the *sample average* $\bar{X} = n^{-1}(X_1 + \dots + X_n)$ is $n^{-2}(n\sigma^2) = \frac{1}{n}\sigma^2$, which tends to zero as n tends to infinity. This observation can be used to prove that, for any $\epsilon > 0$, the probability that $|\bar{X} - \mu|$ is greater than ϵ tends to zero as n tends to infinity. Thus, the sample average “converges in probability” to the mean μ .

This result is called the *weak law of large numbers*. The argument sketched above implicitly assumes that the random variables have finite variance, but this assumption turns out not to be necessary. There is also a *strong law of large numbers*, which states that, with probability 1, the sample average of the first n variables converges to μ as n tends to infinity. As its name suggests, the strong law is stronger than the weak law, in the sense that the weak law can be deduced from the strong law. Notice that these laws make long-term predictions of a statistical kind about the real events that we have chosen to model using probability theory. Moreover, these predictions can be checked experimentally, and the experimental evidence confirms them. This provides a convincing scientific justification for our models.

5 The Normal Distribution and the Central Limit Theorem

As we have seen, for the binomial distribution with parameters n and p , the probability p_k is given by the formula $\binom{n}{k} p^k (1-p)^{n-k}$. If n is large and you plot the points (k, p_k) on a graph, then you will notice that they lie in a bell-shaped curve that has a sharp peak around the mean np . The width of the tall part of the curve has

Note for PUP: from here to the end of this section has changed significantly since the proofreading proof, but the rest of the article is unchanged.

order of magnitude $\sqrt{np(1-p)}$, the standard deviation of the distribution. Let us assume for simplicity that np is an integer, and define a new probability distribution q_k by $q_k = p_{k+np}$. The points (k, q_k) peak at $k = 0$. If you now rescale the graph, compressing horizontally by a factor of $\sqrt{np(1-p)}$ and expanding vertically by the same factor, then the points will all lie close to the graph of

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

This is the density function of a famous distribution known as the *standard normal distribution* on \mathbb{R} . It is also often called the *Gaussian distribution*.

To put this differently, if you toss a biased coin a large number of times, then the number of heads, minus its mean and divided by its standard deviation, is close to a standard normal random variable.

The function $(1/\sqrt{2\pi})e^{-x^2/2}$ occurs in a huge variety of mathematical contexts, from probability theory to FOURIER ANALYSIS [III.27] to quantum mechanics. Why should this be? The answer, as it is for many such questions, is that there are properties that this function has that are shared by no other function.

One such property is *rotational invariance*. Suppose once again that we are throwing a dart at a dartboard and aiming for the bull's-eye. We could model this as the result of adding two independent normal distributions at right angles to each other: one for the x -coordinate and one for the y -coordinate (each having mean 0 and variance 1, say). If we do this, then the two-dimensional “density function” is given by the formula $(1/2\pi)e^{-x^2/2}e^{-y^2/2}$, which can conveniently be written as $(1/2\pi)e^{-r^2/2}$, where r denotes the length of (x, y) . In other words, the density function depends only on the distance from the origin. (This is why it is called “rotationally invariant.”) This very appealing property holds in more dimensions as well. And it turns out to be quite easy to check that $(1/2\pi)e^{-r^2/2}$ is the *only* such function: more precisely, it is the only rotation-invariant density function that makes the coordinates x and y into independent random variables of variance 1. Thus, the normal distribution has a very special symmetry property.

Properties like this go some way toward explaining the ubiquity of the normal distribution in mathematics. However, the normal distribution has an even more remarkable property, which leads to its appearance wherever mathematics is used to model disorder in the real world. The *central limit theorem* states that,

for any sequence of independent and identically distributed random variables X_1, X_2, \dots (with finite mean μ and nonzero finite variance σ^2), we have

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{P}(X_1 + \dots + X_n \leq n\mu + \sqrt{n}\sigma x) \\ = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \end{aligned}$$

for every real number x . The expected value of $X_1 + \dots + X_n$ is $n\mu$ and its standard deviation is $\sqrt{n}\sigma$, so another way of thinking about this is to let $Y_n = (X_1 + \dots + X_n - n\mu)/\sqrt{n}\sigma$. This rescales $X_1 + \dots + X_n$ to have mean 0 and variance 1, and the probability becomes the probability that $Y_n \leq x$. Thus, *whatever* distribution we start with, the limiting distribution of the sum of many independent copies is normal (after appropriate rescaling). Many natural processes can realistically be modeled as accumulations of small independent random effects, and this is why many distributions that one observes, such as the distribution of heights of adults in a given town, have a familiar bell-shaped curve.

A useful application of the central limit theorem is to simplify what look like impossibly complicated calculations. For example, when the parameter n is large, the calculation of binomial probabilities becomes prohibitively complicated. But if X is a binomial random variable, with parameters n and $\frac{1}{2}$, for instance, then we can write X as a sum $Y_1 + \dots + Y_n$, where Y_1, \dots, Y_n are independent Bernoulli random variables with parameter $\frac{1}{2}$. Then, by the central limit theorem,

$$\lim_{n \rightarrow \infty} \mathbb{P}(X \leq \tfrac{1}{2}n + \tfrac{1}{2}\sqrt{n}x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy.$$

III.74 Projective Space

The *real projective plane* can be defined in various ways. One way is to use three *homogeneous coordinates*: a typical point is represented as (x, y, z) , where not all of x , y , and z are equal to 0, with the convention that if λ is a nonzero constant, then (x, y, z) and $(\lambda x, \lambda y, \lambda z)$ are regarded as equal. Notice that for each (x, y, z) the set of all points of the form $(\lambda x, \lambda y, \lambda z)$ is the line through the origin and (x, y, z) , and indeed a more geometrical definition of the real projective plane is that it is the set of all lines in \mathbb{R}^3 that pass through the origin. Each such line meets the unit sphere in exactly two points, which are opposite each other, and a third way of defining the real projective plane is to define opposite points in

the unit sphere to be equivalent and to take the QUOTIENT [I.3 §3.3] of the unit sphere by this EQUIVALENCE RELATION [I.2 §2.3]. A fourth way to define the projective plane is to start with the usual Euclidean plane and to add one “point at infinity” for each possible slope that a line can have. With an appropriate topology, this defines the projective plane as a COMPACTIFICATION [III.9] of the Euclidean plane.

Taking the third definition, a *line* in the projective plane is defined to be a great circle with its opposite points identified. It is then not hard to see that any two lines meet in exactly one point (since any two great circles meet in exactly two opposite points) and that any two points are contained in exactly one line. This property can be used to define much more abstract generalizations of the notion of a projective plane.

Similar definitions hold for other fields besides \mathbb{R} and in higher dimensions. For instance, *complex projective n -space* is the set of all points of the form $(z_1, z_2, \dots, z_{n+1})$, where not every z_i is 0, with $(z_1, z_2, \dots, z_{n+1})$ equivalent to $(\lambda z_1, \lambda z_2, \dots, \lambda z_{n+1})$ if λ is a nonzero complex scalar. This is the set of all “complex lines” in \mathbb{C}^{n+1} that pass through the origin. See SOME FUNDAMENTAL MATHEMATICAL DEFINITIONS [I.3 §6.7] for more details about projective geometry.

III.75 Quadratic Forms

Ben Green

A quadratic form is a homogeneous polynomial of degree 2 in some finite set of unknowns x_1, x_2, \dots, x_n : an example is $q(x_1, x_2, x_3) = x_1^2 - 3x_1x_2 + 4x_3^2$. Here, the coefficients 1, -3, and 4 are integers, but the idea generalizes straightforwardly from \mathbb{Z} to any ring R . Since linear functions are undeniably important and 2 is the next positive integer after 1, one might expect quadratic forms to be important as well, and indeed they are, in many different branches of mathematics, including linear algebra itself.

Here are two theorems about quadratic forms.

Theorem 1. *If \mathbf{x} , \mathbf{y} , and \mathbf{z} are three points in \mathbb{R}^d , then the distances between them satisfy the triangle inequality*

$$|\mathbf{x} - \mathbf{z}| \leq |\mathbf{x} - \mathbf{y}| + |\mathbf{y} - \mathbf{z}|.$$

Theorem 2. *An odd prime p can be written as the sum of two squares if and only if it leaves remainder 1 on division by 4.*

It is not at first sight clear why theorem 1 has anything to do with quadratic forms. The reason is that the square of the *Euclidean distance*

$$|\mathbf{x}| = \sqrt{x_1^2 + \dots + x_d^2}$$

is a quadratic form over the real numbers \mathbb{R} (here, the x_i are the coordinates of \mathbf{x}). This form is derived from the *inner product*

$$\langle \mathbf{x}, \mathbf{y} \rangle = x_1y_1 + \dots + x_dy_d$$

by taking $|\mathbf{x}|^2$ to be $\langle \mathbf{x}, \mathbf{x} \rangle$. The inner product satisfies the relations

- (i) $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ for all $\mathbf{x} \in \mathbb{R}^d$, with equality if and only if $\mathbf{x} = 0$.
- (ii) $\langle \mathbf{x}, \mathbf{y} + \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{x}, \mathbf{z} \rangle$ for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^d$.
- (iii) $\langle \lambda \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, \lambda \mathbf{y} \rangle = \lambda \langle \mathbf{x}, \mathbf{y} \rangle$ for all $\lambda \in \mathbb{R}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.
- (iv) $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

More generally, any function $\phi(\mathbf{x}, \mathbf{y})$ that satisfies these relations is called an inner product. The triangle inequality is a consequence of arguably the most important inequality in mathematics, the CAUCHY-SCHWARZ INEQUALITY [V.22]

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq |\mathbf{x}| |\mathbf{y}|.$$

Not all quadratic forms on \mathbb{R}^d come from inner products, but they do all come from symmetric bilinear forms $g : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$. These are functions of two variables that satisfy all the axioms of an inner product except possibly (i), the positivity criterion. Given a quadratic form $q(\mathbf{x}) = g(\mathbf{x}, \mathbf{x})$, one may recover g using the *polarization identity*

$$g(\mathbf{x}, \mathbf{y}) = \frac{1}{2}(q(\mathbf{x} + \mathbf{y}) - q(\mathbf{x}) - q(\mathbf{y})).$$

This correspondence between quadratic forms and symmetric bilinear forms works just as well when \mathbb{R} is replaced by any field k , except that there are some serious technical issues when k has characteristic two (due to the presence of the fraction $\frac{1}{2}$ in the above formula). In linear algebra one often *defines* quadratic forms by first discussing symmetric bilinear forms. The advantage of this more abstract approach over the concrete definition we gave at the beginning is that it is not necessary to specify a basis for \mathbb{R}^d .

If one makes a good choice of basis, then the quadratic form can be made to look particularly pleasant: we may always choose a basis in such a way that

$$q(\mathbf{x}) = x_1^2 + \dots + x_s^2 - x_{s+1}^2 - \dots - x_t^2$$

for some s and t satisfying $0 \leq s \leq t \leq d$. Here x_1, \dots, x_t are the coefficients of x with respect to the basis we have carefully chosen. The quantity $s - t$ is called the *signature* of the form. When $s = d$ (as is the case for the form defining the Euclidean distance) the form is said to be *positive definite*. Forms that are not positive definite occur very commonly. For example, the form $x^2 + y^2 + z^2 - t^2$ is used to define MINKOWSKI SPACE [I.3 §6.8], which plays a key role in special relativity.

We turn now to examples of quadratic forms in number theory, beginning with two very famous theorems about quadratic forms over the integers \mathbb{Z} . The first is theorem 2, mentioned at the start of the article. It is due to FERMAT [VI.12]. There are many related results for other binary quadratic forms such as $x^2 + 2y^2$ and $x^2 + 3y^2$. In general, however, the question of which primes are represented by $x^2 + ny^2$ is extremely subtle and interesting, and leads one to CLASS FIELD THEORY [V.30].

In 1770 LAGRANGE [VI.22] showed that every number n can be written as a sum of four squares. In fact, the number of such representations of n , $r_4(n)$, is given by the formula

$$r_4(n) = \sum_{\substack{d|n \\ 4 \nmid d}} d.$$

This formula can be explained using the theory of MODULAR FORMS [III.61], one of the most important topics in number theory. Indeed, the generating series

$$f(z) = \sum_{n=0}^{\infty} r_4(n) e^{2\pi i n z}$$

is a *theta series*, as a result of which it satisfies certain transformations that identify it as a modular form.

A remarkable theorem of Conway and Schneeberger states that if a quadratic form $a_1x_1^2 + a_2x_2^2 + a_3x_3^2 + a_4x_4^2$ with $a_1, \dots, a_4 \in \mathbb{N}$ represents all the positive integers less than or equal to 15, then it represents *all* positive integers. RAMANUJAN [VI.82] listed fifty-five such forms; actually, one of his forms did not represent 15, but the remaining fifty-four forms constitute the complete list. For example, every positive integer can be written as $x_1^2 + 2x_2^2 + 4x_3^2 + 13x_4^2$.

Quadratic forms in three variables are more difficult to treat. GAUSS [VI.26] proved that $n = x_1^2 + x_2^2 + x_3^2$ if and only if n does not have the form $4^t(8k+7)$ for integers t and k . It is still not known exactly which integers can be written as $x_1^2 + x_2^2 + 10x_3^2$ (this is known as *Ramanujan's ternary form*).

From the point of view of prime number theory, quadratic forms in *one* variable are the hardest to understand. For example, are there infinitely many primes of the form $x^2 + 1$?

Let us mention one final topic, where quadratic forms over \mathbb{R} are studied but where the unknowns x_1, \dots, x_n are replaced by integers. In particular, let us mention a beautiful result of Margulis, which confirmed a conjecture of Oppenheim. One instance of the result is the following: for any $\epsilon > 0$, one may find integers x_1, x_2 , and x_3 such that

$$0 < |x_1^2 + x_2^2\sqrt{2} - x_3^2\sqrt{3}| < \epsilon.$$

The proof uses techniques from ERGODIC THEORY [V.11], which in related contexts are proving very influential at the forefront of research today. No explicit bounds are known on how large x_1, x_2 , and x_3 need to be.

III.76 Quantum Computation

A quantum computer is a theoretical device that makes use of the phenomenon of “superposition” in quantum mechanics to carry out certain computations in a way that is fundamentally different from any known classical methods, and in a few important cases remarkably efficient. In classical physics, if there is some property that a particle could have, then either it has it or it does not. But according to quantum mechanics, it can exist in a sort of indeterminate state that is a linear combination of several states, in some of which it might have the property in question and in others not. The coefficients in this linear combination are called *probability amplitudes*: the modulus squared of the coefficient associated with a state tells you the probability of finding that the particle is in that state if you do a measurement.

Exactly what happens when you take a measurement is puzzling, and the subject of much debate among physicists and philosophers. Fortunately, however, one can understand quantum computation without solving the measurement problem, as it is called: indeed, one can get away with not understanding quantum mechanics at all. (Similarly, and for similar reasons, one could in principle do significant work in theoretical computer science without having the slightest idea what a transistor is or how it works.)

To understand quantum computation it is helpful to look at two other models of computation. The notion of a *classical* computation is a mathematical distillation of

what actually goes on inside your computer. The “state” of a computer at any one time is modeled by an n -bit string: that is, a sequence of 0s and 1s of length n . Let us write σ for a typical string and $\sigma_1, \sigma_2, \dots, \sigma_n$ for the bits that make it up. A “computation” is a sequence of very simple operations performed on the initial string. For example, one operation might be to choose three numbers i, j , and k , all less than n , and change the k th bit σ_k of the current state σ to 1 if $\sigma_i = \sigma_j = 1$ and to 0 otherwise. What makes an operation such as this “simple” is that it is *local* in character: what it does to σ depends on, and affects, just a bounded number of bits of σ (in this case it depends on two bits and affects one). The “state space” of a classical computer, in this model, is the set $\{0, 1\}^n$ of all possible n -bit strings, which we shall denote by Q_n .

After a certain number of stages, we declare the computation to have finished. At this point we perform a simple sequence of “measurements” on the final state, which consist in looking at the bits of the string we have ended up with. If our problem is a “decision problem,” then we will typically organize the computation so that all we need to look at is a single bit: if it is 0 then the answer is no and if it is 1 then the answer is yes.

If the ideas of the last two paragraphs are unfamiliar to you, then you are strongly advised to read the first few sections of COMPUTATIONAL COMPLEXITY [IV.21] before continuing with this article.

The next model we shall consider is *probabilistic computation*. This is just like classical computation except that at each stage we are allowed to toss a (possibly biased) coin and let the simple operation we perform depend on the outcome of the toss. For instance, we might again choose three numbers i, j , and k , but this time proceed as follows: with probability $\frac{2}{3}$ we perform the operation described earlier, and with probability $\frac{1}{3}$ we change σ_k to $1 - \sigma_k$. Remarkably, introducing randomness into algorithms can be extremely helpful. (Equally remarkably, there are strong theoretical reasons for believing that all algorithms that use randomness can in fact be “derandomized.” See [IV.21 §7.1] for details.)

Suppose that we allow our randomized probabilistic computation to run for k steps and that we do not examine the result. How should we model the current state of the computer? We could use exactly the same definition as in the classical case—a state is an n -bit string—and simply say that the computation is in a state that we cannot know until we do a measurement. But the state of the computer is not a complete mystery:

for each n -bit string σ there will be some probability p_σ that the state is σ . In other words, it is better to think of the state of the computer as a PROBABILITY DISTRIBUTION [III.73] on Q_n . This probability distribution will depend on the initial string, and therefore it can in principle give us useful information about that string.

Here is how to use a randomized computation to solve a decision problem. Let us write $P(\sigma)$ for the probability that a certain bit (without loss of generality the first) is 1 at the end of the computation, when the initial string is σ . Suppose we can arrange for $P(\sigma)$ to be at least a for all strings σ for which the answer is yes, and at most some smaller number b for all strings σ for which the answer is no. Let c be the average of a and b . Now run the computation m times for some large m . With very high probability, if the answer is yes then when we have finished the first bit will have been 1 more than cm times, and if the answer is no then it will have been 1 fewer than cm times. So we can solve the decision problem, not with certainty, but at least with a negligibly small chance of error.

The “state space” of a probabilistic computer consists of all possible probability distributions on Q_n , or equivalently all possible functions $p : Q_n \rightarrow [0, 1]$ such that $\sum_{\sigma \in Q_n} p_\sigma = 1$. The state space of a *quantum* computer also consists of functions defined on Q_n , but there are two differences. First, they can take complex as well as real values. Second, if $\lambda : Q_n \rightarrow \mathbb{C}$ is a state, then the requirement on the size of λ is that $\sum_{\sigma \in Q_n} |\lambda|^2 = 1$. In other words, λ is a unit vector in the HILBERT SPACE [III.37] $\ell_2(Q_n, \mathbb{C})$ rather than a nonnegative unit vector in the BANACH SPACE [III.64] $\ell_1(Q_n, \mathbb{R})$. The scalars λ_σ are the probability amplitudes mentioned earlier. We shall explain what this means later.

Among the possible states of a quantum computer are the “basis states,” which are the functions that take the value 1 at one string and 0 everywhere else. It is customary to use Dirac’s “bra” and “ket” notation for these, writing $|\sigma\rangle$ if the string in question is σ . Other “pure states” are then linear combinations of these, and Dirac’s notation is again used. For instance, if $n = 5$, then one fairly simple state that the computer could be in is $|\psi\rangle = (1/\sqrt{2})|01101\rangle + (i/\sqrt{2})|11001\rangle$.

To get from one state to another, we again apply “local” operations, but adapted to the new, Hilbert space context. Suppose first that we have a basis state $|\sigma\rangle$. Again we look at a very small number of bits. If, for instance, we look at three bits, at i, j , and k , then there

PUP: Tim would like to keep ‘consist in’ here. OK?

PUP: Tim confirms that readers will understand this.

are eight possibilities for the triple $\tau = (\sigma_1, \sigma_2, \sigma_3)$, which we could think of as the basis states in a much smaller state space: the space of all functions $\mu : Q_3 \rightarrow \mathbb{C}$ such that $\sum_{\tau \in Q_3} |\mu_\tau|^2 = 1$. The obvious operations that take unit vectors to unit vectors in a complex Hilbert space are the UNITARY MAPS [III.52 §3.1], and these are indeed what are used.

Let us illustrate this with an example. Suppose that $n = 5$, and that i, j , and k are 1, 2, and 4. One possible operation on these three bits would send $|000\rangle$ to $(|000\rangle + i|111\rangle)/\sqrt{2}$ and $|111\rangle$ to $(i|000\rangle + |111\rangle)/\sqrt{2}$, leaving all other three-bit sequences as they are. If our initial basis state is $|01000\rangle$, then the first, second, and fourth bits are in the state $|000\rangle$, so the resulting state at the end of the operation would be $(|01000\rangle + i|11110\rangle)/\sqrt{2}$.

Now that we have explained what a basic operation does to a basis state, we have in fact explained what it does in general, since the basis states form a basis of the state space. In other words, if you start with a linear combination (or superposition) of basis states, you apply the operation described above to each basis state and take the corresponding linear combination of the results.

Thus, an elementary operation of quantum computation consists in acting on the state space by means of a very special sort of unitary map. If the operation is on k bits (where k is typically very small indeed), then the matrix of this map will be a diagonal sum of 2^{n-k} copies of the $2^k \times 2^k$ unitary matrix used to manipulate those k bits (if the basis elements are appropriately ordered). A quantum computation is a sequence of these elementary operations.

Measuring the result of a quantum computation is more mysterious. The basic idea is simple: we do a certain number of elementary operations and then look at one of the bits of the resulting state. But what does this mean, when the state is not a basis state but rather a superposition of such states? The answer is that when we “measure” the r th bit of the output, we are doing a probabilistic process that is somewhat different from the measurement of a probabilistic computation: if the output state is $\sum_{\sigma \in Q_n} \lambda_\sigma |\sigma\rangle$, then the probability that we observe 1 is the sum of all $|\lambda_\sigma|^2$ such that the k th bit of σ is 1, and the probability that we observe 0 is the same sum but over those σ for which the k th bit is 0. This is why the numbers λ_σ are called probability amplitudes. In order to get a useful answer from a quantum computation, one runs it several times, just as with a probabilistic computation.

Note the following two important differences between a quantum computation and a probabilistic computation. We described the state of a probabilistic computation as a probability distribution on Q_n , which one could also call a convex combination of basis states. But this probability distribution is not telling us what is in the computer: that is a basis state. Rather, it is describing our *knowledge* about what is in the computer. By contrast, the state of a quantum computer *really* is a unit vector in a 2^n -dimensional Hilbert space. So in a certain sense a huge amount of computation can go on in parallel: this is what gives quantum computation its power. Although we cannot know much about the computation, since a single measurement causes it to “collapse,” we can hope to organize it so that different parts of it “interfere” with each other. This “interference” is related to the second main difference, which is the fact that we deal with probability amplitudes rather than probabilities. Roughly speaking, a quantum computation can “split up” and “reassemble itself,” whereas once a probabilistic computation splits up it stays split up. Crucial to the reassembly process in a quantum computation is *cancellation* of probability amplitudes: to give an extreme example, if you multiply a typical unitary matrix by its inverse, then there is a huge amount of cancellation to get all the off-diagonal entries of the resulting matrix to be zero.

All this raises two obvious questions: what are quantum computers good for, and can they actually be built? It turns out that a quantum computer can carry out classical and probabilistic computations, so the first question is asking whether they can do anything further.¹ One might think so, since the state space is so much bigger than it is for a classical computation (it is 2^n dimensional rather than merely n dimensional), and the reassembly process means that we can potentially afford to visit remote parts of the state space, where all coefficients might be of very similar (and small) magnitudes, and come back again to a state where a useful measurement can be made. However, the very vastness of this space means that most states are completely inaccessible unless one is prepared to use a vast number of basic operations. Additionally, it is important that at the end of the computation the output should not be a “typical” state, since only very special states give rise to useful measurements.

1. It is also possible to simulate a quantum computation classically, but it would take an absurdly long time to do so: quantum computers cannot calculate noncomputable functions, but they may be far more efficient at calculating some computable ones.

These arguments show that if a quantum computation is to be useful, then it will have to be very carefully (and cleverly) organized. However, there is a spectacular example of just such a computation: Peter Shor's use of a quantum computer to calculate FAST FOURIER TRANSFORMS [III.26] extremely rapidly. The fast Fourier transform has a symmetry that allows the calculation to be split up and carried out "in parallel" (it might be better to say "in superposition") in a way that is ideally suited to a quantum computer. A super-fast Fourier transform can then be used to solve (by classical methods) some famous computational problems, such as the discrete logarithm problem and the factorization of large integers. The latter can then be used to break a public-key cryptosystem, the encryption method that lies at the heart of modern computer security. (See MATHEMATICS AND CRYPTOGRAPHY [VII.7 §5] and COMPUTATIONAL NUMBER THEORY [IV.5 §3] for further discussion of these problems.)

Can a machine be built that would actually be able to do this? There are formidable problems to overcome, arising from a phenomenon in quantum mechanics known as "decoherence," which makes it very hard to stop a complicated state from "collapsing" to a simpler one that is no longer of use. Some progress has been made, but it is too early to say whether, or when, a quantum computer will be built that can factorize large numbers quickly.

Nevertheless, the theoretical challenges raised by the notion of a quantum computer are fascinating. Perhaps the most interesting one is very simple: find an application of quantum computers that is significantly different from the few that have already been found. The fact that quantum computers can factorize large numbers is strong evidence that they are more powerful, but it would be good to have a better understanding of why. (It is known that quantum computers are better for some other uses, such as COMMUNICATION COMPLEXITY [IV.21 §5.1.4].) Is there a much simpler task that is easy for quantum computers and difficult for classical computers, at least if some well-known plausible hypothesis is true about what classical computers cannot do? Can quantum computers solve NP-COMplete [IV.21 §4] problems? The majority opinion is that they cannot, and indeed the statement that they cannot is becoming another of the many "plausible hypotheses" of complexity theory, but it would be good to have stronger reasons for believing in this statement, such as a proof subject to already-known plausible hypotheses in classical computation.

III.77 Quantum Groups

Shahn Majid

There are at least three different paths that lead to the objects known today as quantum groups. They could be summarized briefly as quantum geometry, quantum symmetry, and self-duality. Any one of them would be a great reason to invent quantum groups and each of them had a role in the development of the modern theory.

1 Quantum Geometry

One of the great discoveries in physics in the twentieth century was that classical mechanics should be replaced by quantum mechanics, in which the space of possible positions and momenta of a particle is replaced by the formulation of position and momentum as mutually noncommuting operators. This noncommutativity underlies Heisenberg's "uncertainty principle," but it also suggests the need for a more general notion of geometry in which coordinates need not commute. One approach to noncommutative geometry is discussed in OPERATOR ALGEBRAS [IV.19 §5]. However, another approach is to note that geometry really grew out of examples such as spheres, tori, and so forth, which are LIE GROUPS [III.50 §1] or objects closely related to Lie groups. If one wants to "quantize" geometry, one should first think about how to generalize basic examples like this: in other words, one should try to define "quantum Lie groups" and associated "quantum" homogeneous spaces.

The first step is to consider geometrical structures not so much in terms of their points but in terms of corresponding *algebras*. For example, the group $SL_2(\mathbb{C})$ is defined as the set of 2×2 matrices $\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$ of complex numbers such that $\alpha\delta - \beta\gamma = 1$. We can think of this as a subset of \mathbb{C}^4 , and indeed not just a subset but a VARIETY [III.97]. The natural class of functions associated with this variety is the set of polynomials in four variables (which are defined on \mathbb{C}^4) restricted to the variety. However, if two polynomials take equal values on the variety, then we identify them. In other words, we take the algebra of polynomials in four variables a, b, c , and d and QUOTIENT [I.3 §3.3] by the IDEAL [III.83 §2] generated by all polynomials of the form $ad - bc - 1$. (This construction is discussed in detail in ARITHMETIC GEOMETRY [IV.6].) Let us call the resulting algebra $\mathbb{C}[SL_2]$.

PUP: I can confirm that this is OK.

We can do the same for any subset $X \subset \mathbb{C}^n$ that is defined by polynomial relations. This gives us a precise one-to-one correspondence between subsets of this type and certain commutative algebras equipped with n generators. Let us write $\mathbb{C}[X]$ for the algebra that corresponds to X . As with many similar constructions (see, for example, the discussion of adjoint maps in DUALITY [III.19]), a suitable map from X to Y gives rise to a map from $\mathbb{C}[Y]$ to $\mathbb{C}[X]$. More precisely, the map ϕ from X to Y has to be polynomial (in a suitable sense) and the resulting map from $\mathbb{C}[Y]$ to $\mathbb{C}[X]$ is an algebra homomorphism ϕ^* that satisfies the formula $\phi^*(p)(x) = p(\phi x)$ for every $x \in X$.

Going back to our example, the set $\mathrm{SL}_2(\mathbb{C})$ has a group structure $\mathrm{SL}_2(\mathbb{C}) \times \mathrm{SL}_2(\mathbb{C}) \rightarrow \mathrm{SL}_2(\mathbb{C})$ defined by the matrix product. The set $\mathrm{SL}_2(\mathbb{C}) \times \mathrm{SL}_2(\mathbb{C})$ is a variety in \mathbb{C}^8 and the matrix product depends in a polynomial way on the entries in the matrices, so we obtain an algebra homomorphism $\Delta : \mathbb{C}[\mathrm{SL}_2] \rightarrow \mathbb{C}[\mathrm{SL}_2] \otimes \mathbb{C}[\mathrm{SL}_2]$, which is known as the *coproduct*. (The algebra $\mathbb{C}[\mathrm{SL}_2] \otimes \mathbb{C}[\mathrm{SL}_2]$ is isomorphic to $\mathbb{C}[\mathrm{SL}_2 \times \mathrm{SL}_2]$.) It turns out that Δ can be expressed by the formula

$$\Delta \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \otimes \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

This formula needs a word or two of explanation: the variables a, b, c , and d are the four generators of the algebra of polynomials in four variables (and hence of its quotient by $ad - bc - 1$), and the right-hand side is a shorthand way of saying that $\Delta a = a \otimes a + b \otimes c$, and so on. Thus, Δ is defined on the generators by a sort of mixture of TENSOR PRODUCTS [III.91] and matrix multiplication.

One can then show that the associativity of matrix multiplication in SL_2 is equivalent to the assertion that $(\Delta \otimes \mathrm{id})\Delta = (\mathrm{id} \otimes \Delta)\Delta$. To understand what these expressions mean, bear in mind that Δ takes elements of $\mathbb{C}[\mathrm{SL}_2]$ to elements of $\mathbb{C}[\mathrm{SL}_2] \otimes \mathbb{C}[\mathrm{SL}_2]$. Thus, when we apply the map $(\Delta \otimes \mathrm{id})\Delta$, for example, we begin by applying Δ , and thereby creating an element of $\mathbb{C}[\mathrm{SL}_2] \otimes \mathbb{C}[\mathrm{SL}_2]$. This element will be a linear combination of elements of the form $p \otimes q$, each of which will then be replaced by $\Delta p \otimes q$.

Similarly, one can express the rest of the group structure of $\mathrm{SL}_2(\mathbb{C})$ equivalently in terms of the algebra $\mathbb{C}[\mathrm{SL}_2]$. There is a *counit* map $\epsilon : \mathbb{C}[\mathrm{SL}_2] \rightarrow k$, which corresponds to the group identity, and an *antipode* map $S : \mathbb{C}[\mathrm{SL}_2] \rightarrow \mathbb{C}[\mathrm{SL}_2]$, which corresponds to the group inversion. The group axioms appear as equivalent properties of these maps, making $\mathbb{C}[\mathrm{SL}_2]$ into a “Hopf alge-

bra” or “quantum group.” The formal definition is as follows.

Definition. A *Hopf algebra* over a field k is a quadruple (H, Δ, ϵ, S) , where

- (i) H is a unital algebra over k ;
- (ii) $\Delta : H \rightarrow H \otimes H$, $\epsilon : H \rightarrow k$ are algebra homomorphisms such that $(\Delta \otimes \mathrm{id})\Delta = (\mathrm{id} \otimes \Delta)\Delta$ and $(\epsilon \otimes \mathrm{id})\Delta = (\mathrm{id} \otimes \epsilon)\Delta = \mathrm{id}$;
- (iii) $S : H \rightarrow H$ is a linear map such that $m(\mathrm{id} \otimes S)\Delta = m(S \otimes \mathrm{id})\Delta = 1\epsilon$, where m is the product operation on H .

There are two great things about this formulation. The first is that the notion of a Hopf algebra makes sense over any field. The second is that nowhere did we demand that H was commutative. Of course, if H is derived from a group, then it certainly is commutative (since multiplying two polynomials is commutative), so if we can find a noncommutative Hopf algebra, then we have obtained a strict generalization of the notion of a group. The great discovery of the past two decades is that there are indeed many natural noncommutative examples.

For example, the quantum group $\mathbb{C}_q[\mathrm{SL}_2]$ is defined as the free associative *noncommutative* algebra on symbols a, b, c , and d modulo the relations

$$\begin{aligned} ba &= qab, & bc &= cb, & ca &= qac, & dc &= qcd, \\ db &= qbd, & da &= ad + (q - q^{-1})bc, & ad - q^{-1}bc &= 1. \end{aligned}$$

This forms a Hopf algebra with Δ given by the same formula as it is for $\mathbb{C}[\mathrm{SL}_2]$ and with suitable maps ϵ and S . Here q is a nonzero element of \mathbb{C} , and as $q \rightarrow 1$ one obtains $\mathbb{C}[\mathrm{SL}_2]$. This example generalizes to canonical examples $\mathbb{C}_q[G]$ for all complex simple Lie groups G .

Much of group theory and Lie group theory can be generalized to quantum groups. For example, Haar integration is a linear map $\int : H \rightarrow k$ that is translation invariant in a certain sense that involves Δ . If it exists, it is unique up to a scalar multiple, and it does indeed exist in most cases of interest, including all finite-dimensional Hopf algebras. Likewise, the notion of a complex of DIFFERENTIAL FORMS [III.16] (Ω, d) makes sense over any algebra H as a proxy for a differential structure. Here, $\Omega = \bigoplus_n \Omega^n$ is required to be an associative algebra generated by $\Omega^0 = H$ and Ω^1 , but one does not assume that it is graded-commutative as in the classical case. When H is a Hopf algebra one can ask that Ω is translation invariant, again in a certain sense that involves the coproduct Δ . In this case both Ω

and its COHOMOLOGY [IV.10 §4] as a complex are super (or graded) quantum groups. The axioms of a (graded) Hopf algebra were originally introduced by Heinz Hopf in 1947 precisely to express the structure of the cohomology ring of a group, so this result brings us back full circle to the origins of the subject. For most quantum groups, including all the $C_q[G]$, one has a natural minimal complex (Ω, d) . Thus, a “quantum group” is not merely a Hopf algebra but has additional structure analogous to that of a Lie group.

There are many other quantum groups that are not related to q -deformations. There are also applications of the theory to finite groups. If G is a finite group, one has a corresponding algebra $k(G)$ of all functions on G with pointwise product and a coproduct $(\Delta f)(g, h) = f(gh)$ for $f \in k(G)$ and $g, h \in G$. Here we identify $k(G) \otimes k(G)$ and $k(G \times G)$, which makes Δf into a function of two variables, and one may check even more simply that this is a Hopf algebra. There can never be an interesting classical differential structure on a finite set, but if we use the methods developed for quantum groups, then we have one or more translation-invariant complexes (Ω^1, d) on any finite group. Applying further parts of the theory of quantum group differential geometry, one finds, for example, that the alternating group A_4 is naturally Ricci-flat, while the symmetric group S_3 naturally has constant CURVATURE [III.13], much like a 3-sphere.

2 Quantum Symmetry

Symmetry in mathematics is usually expressed as the action of a group or Lie algebra of finite or infinitesimal transformations of some structure. If you have a collection of transformations that is closed under inversion and composition, then you necessarily have an ordinary group. So how might one generalize this? The answer is that one begins by observing that a group G can act on several objects at the same time. If a group acts on two objects X and Y , then it also acts on their direct product $X \times Y$, with $g(x, y) = (gx, gy)$. Here we are making implicit use of a diagonal or “duplication” map $\Delta : G \rightarrow G \times G$, which duplicates a group element so that one copy can act on the first object and the other on the second object. In order to generalize this it once again pays to replace the notion of a group G by that of an algebra. This time we use the *group algebra* kG , which is the set of all formal linear combinations $\sum_i \lambda_i g_i$, where the g_i are elements of G and the λ_i are scalars from the field k . The elements

of G (considered as particularly simple linear combinations of this kind) form a basis of kG and we multiply them as we would in G itself. One then extends this definition to products of more general linear combinations in the obvious way. We also extend Δ linearly from $\Delta g = g \otimes g$ on the basis elements to a map from kG to $kG \otimes kG$. Together with some associated maps ϵ and S , this makes kG into a Hopf algebra. Note that this is a completely different use of the coproduct from the one in the previous section, since the group product has already gone into the algebra. One has a similar story for the “enveloping algebra” $U(\mathfrak{g})$ associated with any Lie algebra \mathfrak{g} ; this is generated by a basis of \mathfrak{g} with certain relations and becomes a Hopf algebra with the coproduct $\Delta \xi = \xi \otimes 1 + 1 \otimes \xi$ “sharing out” an element $\xi \in \mathfrak{g}$ for the purposes of acting on a tensor product of objects on which \mathfrak{g} acts.

Extrapolating from these two examples, a general “quantum symmetry” means an algebra H equipped with further structure Δ that allows one to form a tensor product $V \otimes W$ of any two representations V, W of the algebra in an associative manner. An element $h \in H$ acts as $h(v \otimes w) = (\Delta h)(v \otimes w)$, where one part of Δh acts on $v \in V$ and another part on $w \in W$. This is a second route to the Hopf algebra axioms we gave in the previous section.

Note that, in the examples just given, Δ has had a symmetric output. As a consequence, if V and W are representations of a group or Lie algebra, then $V \otimes W$ and $W \otimes V$ are isomorphic via the obvious map that takes $v \otimes w$ to $w \otimes v$. In general, however, $V \otimes W$ and $W \otimes V$ may be unrelated, so it is now the tensor product that is being made noncommutative. In nice examples it may be the case that $V \otimes W \cong W \otimes V$, but not necessarily by the obvious map. Instead, there may be a nontrivial isomorphism for every pair V, W , which may nevertheless obey some reasonable conditions. This happens for a large class of examples, denoted by $U_q(\mathfrak{g})$ and associated with all complex simple Lie algebras. For these examples, the isomorphism obeys the braid or Yang-Baxter relations among any three representations (see BRAID GROUPS [III.4]). As a result, these quantum groups lead to KNOT AND 3-MANIFOLD INVARIANTS [III.46] (the Jones knot invariant comes from the example $U_q(\mathfrak{sl}_2)$, where \mathfrak{sl}_2 is the Lie algebra of the group $SL_2(\mathbb{C})$). The parameter q can usefully be regarded here as a formal variable, and these examples can be thought of as some kind of deformation of the classical enveloping algebras $U(\mathfrak{g})$. They arose originally in work of Drin-

feld and of Jimbo in the theory of quantum integrable systems.

3 Self-duality

A third point of view is that Hopf algebras are the next simplest CATEGORY [III.8] after Abelian groups of structures that admit a FOURIER TRANSFORM [III.27]. It is not immediately obvious, but the axioms (i)–(iii) in the definition we gave earlier have a certain symmetry. One can write out the requirement (i) of a unital algebra H in terms of linear maps $m : H \otimes H \rightarrow k$ and $\eta : k \rightarrow H$ (here η specifies the identity element of H as the image of $1 \in k$) that have to obey some straightforward commutative diagrams. If you reverse all the arrows in these diagrams, then you have the axioms displayed in (ii), obtaining what could be called a “coalgebra.” The requirement that the coalgebra structures Δ and ϵ are algebra maps is given by a collection of diagrams that is invariant under arrow reversal. Finally, the axioms in (iii), as commutative diagrams, are invariant under arrow reversal in the above sense.

Thus, the axioms of a Hopf algebra have the special property of being symmetric under arrow reversal. A practical consequence is that if H is a finite-dimensional Hopf algebra, then so is H^* , with all structure maps defined as the adjoints of those of H (which necessarily reverses arrows). In the infinite-dimensional case one needs a suitable topological dual, or one can just speak of two Hopf algebras as dually paired to each other. For instance, $\mathbb{C}_q[\mathrm{SL}_2]$ and $U_q(\mathfrak{sl}_2)$ above are dually paired, while if G is finite then $(kG)^* = k(G)$, the Hopf algebra of functions on G .

As an application, let H be finite dimensional with basis $\{e_a\}$, let H^* have a dual basis $\{f^a\}$, and let \int denote a right-translation-invariant integral on H . The Fourier transform $\mathcal{F} : H \rightarrow H^*$ is defined as

$$\mathcal{F}(h) = \sum_a \left(\int e_a h \right) f^a$$

and has many remarkable properties. A special case is a Fourier transform $\mathcal{F} : k(G) \rightarrow kG$ for any finite group G , which does not have to be Abelian. If G happens to be Abelian, then $kG \cong k(\hat{G})$, where \hat{G} is the group of characters, and we recover the usual Fourier transform for finite Abelian groups. The point is that in the non-Abelian case, kG is not commutative and hence not the algebra of functions on any usual “Fourier dual” space.

This point of view is responsible for the second main class of genuine quantum groups to have been

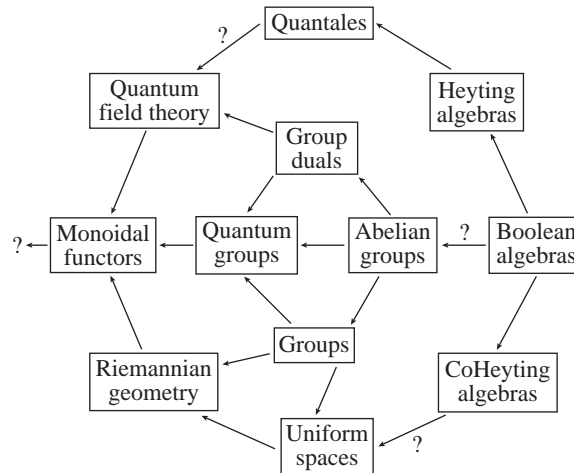


Figure 1 Putting quantum groups in context. Self-dual categories are shown on the horizontal axis.

discovered, namely the “bicrossproduct” ones of self-dual form. They are simultaneously “coordinate” and “symmetry” algebras, and are truly connected with quantum mechanics. An example, which is written $\mathbb{C}[\mathbb{R}^3 \rtimes \mathbb{R}]_\lambda \curvearrowright U(\mathfrak{so}(1,3))$, is the so-called *Poincaré quantum group* of a certain noncommutative spacetime with coordinates x, y, z, t , where t does not commute with the other variables. This quantum group can also be interpreted as the quantization of a particle moving in a curved geometry with black-hole-like features. In essence, the self-duality of quantum groups provides a paradigm for “toy models” of the unification of gravity (as spacetime geometry) and quantum theory.

This is part of a wider picture indicated in figure 1. A category of objects with a coherent notion of “tensor product” is called a *monoidal* (or *tensor*) *category*, and we have seen that this is the case for representations of quantum groups. There, one also has a “forgetful functor” to the category of vector spaces, which forgets the quantum group action. This embeds quantum groups into the next most general self-dual category (in a representation-theoretic sense), namely that of functors between monoidal categories. Over on the right, I have included Boolean algebras as primitive structures with (de Morgan) duality. However, the connection between duality here and the other dualities is speculative.

Further Reading

Majid, S. 2002. *A Quantum Groups Primer*. London Mathematical Society Lecture Notes, volume 292. Cambridge:

T&T note: check two symbols here before CRC.

PUP: ‘Quantales’ is indeed OK in the figure.

Cambridge University Press.

III.78 Quaternions, Octonions, and Normed Division Algebras

Mathematics took a leap forward in sophistication with the introduction of the COMPLEX NUMBERS [I.3 §1.5]. To define these, one suspends one's disbelief, introduces a new number i , and declares that $i^2 = -1$. A typical complex number is of the form $a + ib$, and the arithmetic of complex numbers is easy to deduce from the normal rules of arithmetic for real numbers. For example, to calculate the product of $1 + 2i$ and $2 + i$ one simply expands some brackets:

$$(1 + 2i)(2 + i) = 2 + 5i + 2i^2 = 5i,$$

the last equality following from the fact that $i^2 = -1$. One of the great advantages of the complex numbers is that, if complex roots are allowed, every polynomial can be factorized into linear factors: this is the famous FUNDAMENTAL THEOREM OF ALGEBRA [V.15].

Another way to define a complex number is to say that it is a pair of real numbers. That is, instead of writing $a + ib$ one writes simply (a, b) . To add two complex numbers is simple, and exactly what one does when adding two vectors: $(a, b) + (c, d) = (a + c, b + d)$. However, it is less obvious how to multiply: the product of (a, b) and (c, d) is $(ac - bd, ad + bc)$, which seems an odd definition unless one goes back to thinking of (a, b) and (c, d) as $a + ib$ and $c + id$.

Nevertheless, the second definition draws our attention to the fact that the complex numbers are formed out of the two-dimensional VECTOR SPACE [I.3 §2.3] \mathbb{R}^2 with a carefully chosen definition of multiplication. This immediately raises a question: could we do the same for higher-dimensional spaces?

As it stands, this question is not wholly precise, since we have not been clear about what "the same" means. To make it precise, we must ask what properties this multiplication should have. So let us return to \mathbb{R}^2 and think about why it would be a bad idea to define the product of (a, b) and (c, d) in a simple-minded way as (ac, bd) . Of course, part of the reason is that the product of $a + ib$ and $c + id$ is not $ac + ibd$, but why should we not also be interested in other ways of multiplying vectors in \mathbb{R}^2 ?

The trouble with this alternative definition is that it allows *zero divisors*, that is, pairs of nonzero numbers that multiply together to give zero. For example,

it gives us $(1, 0)(0, 1) = (0, 0)$. If we have zero divisors, then we cannot have multiplicative inverses, since if every nonzero number in a number system has a multiplicative inverse, and if $xy = 0$, then either $x = 0$ or $y = x^{-1}xy = x^{-1}0 = 0$. And if we do not have multiplicative inverses, then we cannot define a useful notion of division.

Let us return then to the usual definition of the complex numbers and try to think how we can go beyond it. One way we might try to "do the same" as we did before is to do to the complex numbers what we did to the real numbers. That is, why not define a "super-complex" number to be an ordered pair (z, w) of complex numbers? Since we still want to have a vector space, we will continue to define the sum of (z, w) and (u, v) to be $(z + u, w + v)$, but we need to think about the best way of defining their product. An obvious guess is to use precisely the expression that worked before, namely $(zu - wv, zv + wu)$. But if we do that, then the product of $(1, i)$ and $(1, -i)$ works out to be $(1 + i^2, i - i) = (0, 0)$, so we have zero divisors.

This example came from the following thought. The *modulus* of a complex number $z = a + ib$, which measures the length of the vector (a, b) , is the real number $|z| = \sqrt{a^2 + b^2}$. This can also be written as $\sqrt{z\bar{z}}$, where \bar{z} is the *complex conjugate* $a - ib$ of z . Now if a and b are allowed to take *complex* values, then there is no reason for $a^2 + b^2$ to be nonnegative, so we may not be able to take its square root. Moreover, if $a^2 + b^2 = 0$ it does not follow that $a = b = 0$. The example above came from taking $a = 1$ and $b = i$ and multiplying the number $(1, i)$ by its "conjugate" $(1, -i)$.

There is, nevertheless, a natural way to define the modulus of a pair (z, w) that works even when z and w are complex numbers. The number $|z|^2 + |w|^2$ is guaranteed to be nonnegative, so we can take its square root. Moreover, if $z = a + ib$ and $w = c + id$, then we will obtain the number $(a^2 + b^2 + c^2 + d^2)^{1/2}$, which is the length of the vector (a, b, c, d) .

This observation leads to another: the complex conjugate of a real number is the number itself, so, if we want to "use the same formula" for the complex numbers as we used for the reals, we are free to introduce complex conjugates into that formula. Before we try to do that, let us think about what we might mean by the "conjugate" of a pair (z, w) . We expect $(z, 0)$ to behave like the complex number z , so its conjugate should be $(\bar{z}, 0)$. Similarly, if z and w are real, then the conjugate of (z, w) should be $(z, -w)$. This leaves us with two

reasonable possibilities for a general pair (z, w) : either $(\bar{z}, -\bar{w})$ or $(\bar{z}, -w)$. Let us consider the second of these.

We would like the product of (z, w) and its conjugate, which we are defining as $(\bar{z}, -w)$, to be $(|z|^2 + |w|^2, 0)$. We want to achieve this by introducing complex conjugates into the formula

$$(z, w)(u, v) = (zu - wv, zv + wu).$$

An obvious way of getting the result we want is to take

$$(z, w)(u, v) = (zu - \bar{w}v, \bar{z}v + wu),$$

and this modified formula, it turns out, defines an ASSOCIATIVE BINARY OPERATION [I.2 §2.4] on the set of pairs (z, w) . If you try the other definition of conjugate, you will find that you end up with zero divisors. (A first indication of trouble is that, under the other definition, the pair $(0, i)$ is its own conjugate.)

We have just defined the *quaternions*, a set \mathbb{H} of “numbers” that form a four-dimensional real vector space, or alternatively a two-dimensional complex vector space. (The letter “H” is in honor of William Rowan Hamilton, their discoverer. See HAMILTON [VI.37] for the story of how the discovery was made.) But why should we have wished to do that? This question becomes particularly pressing when we notice that the notion of multiplication that we have defined is not commutative. For example, $(0, 1)(i, 0) = (0, i)$, while $(i, 0)(0, 1) = (0, -i)$.

To answer it, let us take a step back and think about the complex numbers again. The most obvious justification for introducing those is that one can use them to solve all polynomial equations, but that is by no means the only justification. In particular, complex numbers have an important *geometrical* interpretation, as rotations and enlargements. This connection becomes particularly clear if we choose yet another way of writing the complex number $a + ib$, as the matrix $\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$. Multiplication by the complex number $a + ib$ can be thought of as a LINEAR MAP [I.3 §4.2] on the plane \mathbb{R}^2 , and this is the matrix of that linear map. For example, the complex number i corresponds to the matrix $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, which is the matrix of a counterclockwise rotation through $\frac{1}{2}\pi$, and this rotation is exactly what multiplying by i does to the complex plane.

If complex numbers can be thought of as linear maps from \mathbb{R}^2 to \mathbb{R}^2 , then quaternions should have an interpretation as linear maps from \mathbb{C}^2 to \mathbb{C}^2 . And indeed they do. Let us associate with the pair (z, w) the matrix $\begin{pmatrix} z & \bar{w} \\ -w & \bar{z} \end{pmatrix}$. Now let us consider the product of two such

matrices:

$$\begin{pmatrix} z & \bar{w} \\ -w & \bar{z} \end{pmatrix} \begin{pmatrix} u & \bar{v} \\ -v & \bar{u} \end{pmatrix} = \begin{pmatrix} zu - \bar{w}v & z\bar{v} + \bar{w}\bar{u} \\ -\bar{z}v - wu & \bar{z}\bar{u} - w\bar{v} \end{pmatrix}.$$

This is precisely the matrix associated with the pair $(zu - w\bar{v}, z\bar{v} + w\bar{u})$, which is the quaternionic product of (z, w) and (u, v) ! As an immediate corollary, we have a proof of a fact mentioned earlier: that quaternionic multiplication is associative. Why? Because matrix multiplication is associative. (And *that* is true because the composition of functions is associative: see [I.3 §3.2].)

Notice that the DETERMINANT [III.15] of the matrix $\begin{pmatrix} z & \bar{w} \\ -w & \bar{z} \end{pmatrix}$ is $|z|^2 + |w|^2$, so the modulus of the pair (z, w) (which is defined to be $\sqrt{|z|^2 + |w|^2}$) is just the determinant of the associated matrix. This proves that the modulus of the product of two quaternions is the product of their moduli (since the determinant of a product is the product of determinants). Notice also that the adjoint of the matrix (that is, the complex conjugate of the transpose matrix) is $\begin{pmatrix} \bar{z} & -\bar{w} \\ w & z \end{pmatrix}$, which is the matrix associated with the conjugate pair $(\bar{z}, -w)$. Finally, notice that if $|z|^2 + |w|^2 = 1$, then

$$\begin{pmatrix} z & \bar{w} \\ -w & \bar{z} \end{pmatrix} \begin{pmatrix} \bar{z} & -\bar{w} \\ w & z \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

which tells us that the matrix is UNITARY [III.52 §3.1]. Conversely, any unitary 2×2 matrix with determinant 1 can easily be shown to have the form $\begin{pmatrix} z & \bar{w} \\ -w & \bar{z} \end{pmatrix}$. Therefore, the unit quaternions (that is, the quaternions of modulus 1) have a geometrical interpretation: they correspond to the “rotations” of \mathbb{C}^2 (that is, the unitary maps of determinant 1), just as the unit complex numbers correspond to the rotations of \mathbb{R}^2 .

The group of unitary transformations of \mathbb{C}^2 of determinant 1 is an important LIE GROUP [III.50 §1] called the *special unitary group* $SU(2)$. Another important Lie group is the group $SO(3)$, of rotations of \mathbb{R}^3 . Surprisingly, the unit quaternions can be used to describe this group as well. To see this, it is convenient to present the quaternions in another, more conventional, way.

Quaternions, as they are usually introduced, are a system of numbers where -1 has not just one square root but three, called i , j , and k . Once one knows that $i^2 = j^2 = k^2 = -1$, and also that $ij = k$, $jk = i$, and $ki = j$, one has all the information one needs to multiply two quaternions. For example, $ji = jik = -k$. A typical quaternion takes the form $a + ib + jc + kd$, which corresponds to the pair of complex numbers $(a + ic, b + id)$ in our previous way of thinking about quaternions. Now if we want, we can think of this quaternion as a pair

(a, \mathbf{v}) , where a is a real number and \mathbf{v} is the vector (b, c, d) in \mathbb{R}^3 . The product of (a, \mathbf{v}) and (b, \mathbf{w}) then works out to be $(ab - \mathbf{v} \cdot \mathbf{w}, a\mathbf{w} + b\mathbf{v} + \mathbf{v} \wedge \mathbf{w})$, where $\mathbf{v} \cdot \mathbf{w}$ and $\mathbf{v} \wedge \mathbf{w}$ are the scalar and vector products of \mathbf{v} and \mathbf{w} .

If $q = (a, \mathbf{u})$ is a quaternion of modulus 1, then $a^2 + \|\mathbf{u}\|^2 = 1$, so we can write q in the form $(\cos \theta, \mathbf{v} \sin \theta)$ with \mathbf{v} a unit vector. This quaternion corresponds to a counterclockwise rotation R about an axis in direction \mathbf{v} through an angle of 2θ . This angle is not what one might at first expect, and neither is the way the correspondence works. If \mathbf{w} is another vector, we can represent it as the quaternion $(0, \mathbf{w})$. We would now like a neat expression for the quaternion $(0, R\mathbf{w})$; it turns out that $(0, R\mathbf{w}) = q(0, \mathbf{w})q^*$, where q^* is the conjugate $(\cos \theta, -\mathbf{v} \sin \theta)$ of q , which is also its multiplicative inverse, as q has modulus 1. So to do the rotation R , you do not multiply by q but rather you *conjugate* by q . (This is a different meaning of the word “conjugate,” referring to multiplying on one side by q and on the other side by q^{-1} .) Now if q_1 and q_2 are quaternions corresponding to rotations R_1 and R_2 , respectively, then

$$q_2 q_1 (0, \mathbf{w}) q_1^* q_2^* = q_2 q_1 (0, \mathbf{w}) (q_2 q_1)^*,$$

from which it follows that $q_2 q_1$ corresponds to the rotation $R_2 R_1$. This tells us that quaternionic multiplication corresponds to composition of rotations.

The unit quaternions form a group, as we have already seen—it is $SU(2)$. It might appear that we have shown that $SU(2)$ is the same as the group $SO(3)$ of rotations of \mathbb{R}^3 . However, we have not quite done this, because for each rotation of \mathbb{R}^3 there are *two* unit quaternions that give rise to it. The reason is simple: a counterclockwise rotation through θ about a vector \mathbf{v} is the same as a counterclockwise rotation through $-\theta$ about $-\mathbf{v}$. In other words, if q is a unit quaternion, then q and $-q$ give rise to the same rotation of \mathbb{R}^3 . So $SU(2)$ is not isomorphic to $SO(3)$; rather, it is a *double cover* of $SO(3)$. This fact has important ramifications in mathematics and physics. In particular, it lies behind the notion of the “spin” of an elementary particle.

Let us return to the question we were considering earlier: for which n is there a good way of multiplying vectors in \mathbb{R}^n ? We now know that we can do it for $n = 1, 2$, or 4 . When $n = 4$ we had to sacrifice commutativity, but we were amply rewarded for this, since quaternion multiplication gives a very concise way of representing the important groups $SU(2)$ and $SO(3)$. These groups

are not commutative, so it was essential to our success that quaternion multiplication should also not be commutative.

One obvious thing we can do is continue the process that led to the quaternions. That is, we can consider pairs (q, r) of quaternions, and multiply these pairs by the formula

$$(q, r)(s, t) = (qs - r^*t, q^*t + rs).$$

Since the conjugate q^* of a quaternion q is the analogue of the complex conjugate \bar{z} of a complex number z , this is basically the same formula that we used for multiplication of pairs of complex numbers—that is, for quaternions.

However, we need to be careful: multiplication of quaternions is not commutative, so there are in fact many formulas we could write down that would be “basically the same” as the earlier one. Why choose the above one, rather than, say, replacing q^*t by tq^* ?

It turns out that the formula suggested above leads to zero divisors. For example, $(j, i)(l, k)$ works out to be $(0, 0)$. However, the modified formula

$$(q, r)(s, t) = (qs - tr^*, q^*t + sr),$$

which one can discover fairly quickly if one bears in mind that one would like $(q, r)(q^*, -r)$ to work out as $(|q|^2 + |r|^2, 0)$, does produce a useful number system. It is denoted \mathbb{O} and its elements are called the *octonions* (or sometimes the *Cayley numbers*). Unfortunately, multiplication of octonions is not even associative, but it does have two very good properties: every nonzero octonion has a multiplicative inverse, and two nonzero octonions never multiply together to give zero. (Because octonion multiplication is not associative, these two properties are no longer obviously equivalent. However, any subalgebra of the octonions generated by two elements *is* associative, and this is enough to prove the equivalence.)

So now we have number systems when $n = 1, 2, 4$, or 8 . It turns out that these are the *only* dimensions with good notions of multiplication. Of course, “good” has a technical meaning here: matrix multiplication, which is associative but gives zero divisors, is for many purposes “better” than octonion multiplication, which has no zero divisors but is not associative. So let us finish by seeing more precisely what it is that is special about dimensions 1, 2, 4, and 8.

All the number systems constructed above have a notion of size given by a NORM [III.64]. For real and complex numbers z , the norm of z is just its modulus. For

a quaternion or octonion x , it is defined to be $\sqrt{x^*x}$, where x^* is the conjugate of x (a definition that works for real and complex numbers as well). If we write $\|x\|$ for the norm of x , then the norms constructed have the property that $\|xy\| = \|x\| \|y\|$ for every x and y . This property is extremely useful: for example, it tells us that the elements of norm 1 are closed under multiplication, a fact that we used many times when discussing the geometric importance of complex numbers and quaternions.

The feature that distinguishes dimensions 1, 2, 4, and 8 from all other dimensions is that these are the only dimensions for which one can define a norm $\|\cdot\|$ and a notion of multiplication with the following properties.

- (i) There is a multiplicative identity: that is, a number 1 such that $1x = x1 = x$ for every x .
- (ii) Multiplication is *bilinear*, meaning that $x(y+z) = xy + xz$, and $x(ay) = a(xy)$ whenever a is a real number.
- (iii) For any x and y , $\|xy\| = \|x\| \|y\|$ (and therefore there are no zero divisors).

A *normed division algebra* is a vector space \mathbb{R}^n together with a norm and a method of multiplying vectors that satisfy the above properties. So normed division algebras exist only in dimensions 1, 2, 4, and 8. Furthermore, even in these dimensions, \mathbb{R} , \mathbb{C} , \mathbb{H} , and \mathbb{O} are the only examples.

There are various ways to prove this fact, which is known as *Hurwitz's theorem*. Here is a very brief sketch of one of them. The idea is to prove that if a normed division algebra A contains one of the above examples, then either it *is* that example, or it contains the next one in the sequence. So either A is one of \mathbb{R} , \mathbb{C} , \mathbb{H} , and \mathbb{O} or A contains the algebra produced by doing to \mathbb{O} the process we used to construct \mathbb{H} from \mathbb{C} and \mathbb{O} from \mathbb{H} , a process known as the *Cayley-Dickson construction*. However, if one applies the Cayley-Dickson construction to \mathbb{O} , one obtains an algebra with zero divisors.

To see how such an argument might work, let us imagine, for the sake of example, that A contains \mathbb{O} as a proper subalgebra. It turns out that the norm on A must be a *EUCLIDEAN NORM* [III.37]—that is, a norm derived from an inner product. (Roughly speaking, this is because multiplication by an element of norm 1 does not change the norm, which gives A so many symmetries that the norm on A has to be the most symmetric of all, namely Euclidean.) Let us call an element of A *imaginary* if it is orthogonal to the element 1. Then

we can define a conjugation operation on A by taking 1^* to be 1 and x^* to be $-x$ when x is imaginary, and extending linearly. This operation can be shown to have all the properties one would like. In particular, $aa^* = a^*a = \|a\|^2$ for every element a of A . Let us choose a norm-1 element of A that is orthogonal to all of \mathbb{O} and call it i . Then $i^* = -i$, so $1 = i^*i = -i^2$, so $i^2 = -1$. Now take the algebra generated by i and the copy of \mathbb{O} that lies in A . With some algebraic manipulation, one can demonstrate that this consists of elements of the form $x + iy$, with x and y belonging to \mathbb{O} . Moreover, the product of $x + iy$ and $z + iw$ turns out to be $xz - wy^* + i(x^*w + zy)$, which is exactly what the Cayley-Dickson construction gives.

For further details about quaternions and octonions, there are two excellent sources: a discussion by John Baez at <http://math.ucr.edu/home/baez/> octonions and a book, *On Quaternions and Octonions: Their Geometry, Arithmetic, and Symmetry*, by J. H. Conway and D. A. Smith (2003; Wellesley, MA: AK Peters).

III.79 Representations

A *linear representation* of a finite GROUP [I.3 §2.1] G is a way of associating a linear map T_g , from some VECTOR SPACE [I.3 §2.3] V to itself, with each element g of G . Of course, this association must reflect the group structure of G , so T_gT_h should equal T_{gh} , and if e is the identity of G , then T_e should be the identity map on V .

One useful aspect of linear representations is that the dimension of the vector space V may be considerably smaller than the size of G . If this is the case, then the representation packages the information about G in a particularly efficient way. For example, the *ALTERNATING GROUP* [III.70] A_5 , which has sixty elements, is isomorphic to the group of rotational symmetries of an icosahedron, and can therefore be thought of as a group of transformations of \mathbb{R}^3 (or, equivalently, of 3×3 matrices).

A more fundamental reason for representations being useful is that every representation can be decomposed into building blocks known as *irreducible* representations. It turns out that a great deal of information about G can be deduced from a few basic facts about its irreducible representations.

These ideas can be generalized to infinite groups as well, and are particularly important in the case of

LIE GROUPS [III.50 §1]. Since Lie groups have a differentiable structure, the representations of interest are those where the homomorphism $g \mapsto T_g$ reflects this structure (for example, by being differentiable).

Representations are discussed in much greater detail in REPRESENTATION THEORY [IV.12]. See also OPERATOR ALGEBRAS [IV.19 §2].

III.80 Ricci Flow

Terence Tao

Ricci flow is a technique that allows one to take an arbitrary RIEMANNIAN MANIFOLD [I.3 §6.10] and smooth out the geometry of that manifold to make it look more symmetric. It has proven to be a very useful tool in understanding the topology of such manifolds.

Ricci flow can be defined for Riemannian manifolds of any dimension, but for the sake of exposition we restrict ourselves here to two-dimensional manifolds (i.e., surfaces) as they are easy to visualize. From our everyday experience with three-dimensional space \mathbb{R}^3 , we are familiar with many surfaces, such as spheres, cylinders, planes, tori (the shape of the surface of a doughnut), and so forth. This is an *extrinsic* way to think about surfaces: as subsets of a larger *ambient space*, which in this case is three-dimensional Euclidean space. On the other hand, one can think about surfaces in a more abstract *intrinsic* manner: by considering how the points in the surface stand in relation to each other, but not in relation to any external space. (For instance, the Klein bottle makes perfect sense as a surface from an intrinsic viewpoint, but cannot be viewed extrinsically in three-dimensional Euclidean space \mathbb{R}^3 , although it can be viewed extrinsically in four-dimensional Euclidean space \mathbb{R}^4 .) It turns out that the two viewpoints are mostly equivalent to each other, but it will be more convenient here to adopt the intrinsic perspective.

A good example of a surface is the surface of Earth. Extrinsically, this is a subset of a three-dimensional space \mathbb{R}^3 . But we can also view this surface two dimensionally by using an *atlas*: a collection of *maps* or *charts* that describe various regions of this surface by identifying them with a subset of a two-dimensional plane. As long as we have enough charts to cover the original surface, this atlas is sufficient to describe the surface. This way of thinking of a surface is not completely intrinsic, because there is more than one atlas that one could associate with this surface, and they may differ in various minor ways. For instance, in one atlas the city

of Los Angeles might be on the boundary of one of the charts, whereas in another atlas it might be in the interior of every chart that it appears in. However, there are many facts one can deduce from an atlas that do not depend on the choice of atlas; for instance, using any accurate atlas of Earth one can see that it is impossible to travel from Los Angeles to Sydney without crossing at least one ocean. If a fact regarding a surface does not depend on which atlas one uses, we say that it is *intrinsic* or *coordinate-independent*. It will turn out that Ricci flow is an intrinsic flow on surfaces; it can be defined without any knowledge of charts or of some external space.

We have informally described the mathematical concept of a surface, or two-dimensional manifold. But to describe Ricci flow we need the more sophisticated concept of a *Riemannian surface* (or two-dimensional Riemannian manifold). This is a surface M with an additional (intrinsic) object, a *Riemannian metric* g , which specifies the distance $d(x, y)$ between any two points x, y on the surface. This metric allows one to define the angle $\angle y_1, y_2$ that any two curves y_1, y_2 on the surface make where they intersect; for instance, the Earth's equator intersects any longitude at right angles. And it can also be used to define the area $|A|$ of any given set A on the surface (e.g., the area of Australia). There are a number of properties that these concepts of distance, angle, and area have to satisfy, but the most important property can be stated informally as follows: *the geometry of a Riemannian surface has to be very close to the geometry of the Euclidean plane at small length scales*.

To give an example of what the above statement means, take any point x in the surface M , and pick any positive radius r . Because the Riemannian metric g specifies a notion of distance, we can define the *disk* $B(x, r)$ of radius r centered at x to be the set of all points y whose distance $d(x, y)$ to x is less than r . Because the Riemannian metric g defines a notion of area, we can then discuss the area of this disk $B(x, r)$. In the Euclidean plane, this area would of course be πr^2 . In a Riemannian surface, this need not be the case: for instance, the total area of the surface of Earth (and hence of all disks within this surface) is finite, even though πr^2 can be arbitrarily large as r goes to infinity. However, we do require that, when r is very small, the area of the disk $B(x, r)$ becomes increasingly close to πr^2 ; more precisely, we require that the ratio between the area and πr^2 converges to 1 in the limit as r tends to 0.

This brings us to the notion of *scalar curvature* $R(x)$. In some cases, such as on the sphere, the area $|B(x, r)|$ of a small disk $B(x, r)$ is actually a little bit smaller than πr^2 ; when this is the case, we say that the surface has *positive scalar curvature* at x . In some other cases, such as on a saddle, the area $|B(x, r)|$ of a small disk $B(x, r)$ is a bit larger than πr^2 ; then we say that the surface has *negative scalar curvature* at x . In other cases again, such as on a cylinder, the area $|B(x, r)|$ of a small disk $B(x, r)$ is equal (or very nearly equal) to πr^2 ; in this case we say the surface has *vanishing scalar curvature* at x . (This is despite the cylinder being “curved” when viewed extrinsically as a subset of three-dimensional space.) Note that on a complicated surface it is perfectly possible to have positive scalar curvature at some points of the surface and negative or vanishing scalar curvature at other points. The scalar curvature $R(x)$ at any given point x can be defined more precisely by the formula

$$R(x) = \lim_{r \rightarrow 0} \frac{\pi r^2 - |B(x, r)|}{\pi r^4 / 24}.$$

(For surfaces in an external space, this intrinsic concept of scalar curvature is almost identical to the extrinsic concept of *Gauss curvature*, which we will not discuss here.)

One can refine this notion to that of *Ricci curvature* $\text{Ric}(x)(v, v)$. Consider now an angular sector $A(x, r, \theta, v)$ inside a small disk $B(x, r)$ of small angular aperture θ (measured in radians) about some direction v (a unit vector) emanating from x . This sector is well-defined, basically because the Riemannian metric gives us the appropriate notions of distance and angle. In Euclidean space, the area $|A(x, r, \theta, v)|$ of this sector is $\frac{1}{2}\omega r^2$. But on a surface, the area $|A(x, r, \theta, v)|$ might be slightly less (respectively, slightly more) than $\frac{1}{2}\omega r^2$. In these cases we say that the surface has positive (respectively, negative) Ricci curvature at x in the direction v . More precisely, we have

$$\text{Ric}(x)(v, v) = \lim_{r \rightarrow 0} \lim_{\omega \rightarrow 0} \frac{\frac{1}{2}\omega r^2 - |A(x, r, \omega, v)|}{\omega r^4 / 24}.$$

Now it turns out that for surfaces, this more complicated notion of curvature is in fact equal to half the scalar curvature: $\text{Ric}(x)(v, v) = \frac{1}{2}R(x)$. In particular, the direction v plays no role in Ricci curvature in two dimensions. However, it is possible to extend all of the above concepts to other dimensions. (For instance, to define scalar and Ricci curvature for three-dimensional manifolds, one would use balls and solid sectors instead of disks and angular sectors, as well as

making other necessary adjustments, such as replacing the expression πr^2 with $\frac{4}{3}\pi r^3$.) In higher dimensions it turns out that the Ricci curvature is more complicated than the scalar curvature. For instance, in three dimensions it is possible for a point x to have positive Ricci curvature in one direction but negative Ricci curvature in another; intuitively, this means that narrow sectors in the former direction “curve inward,” whereas narrow sectors in the latter direction “curve outward.”

Now we can describe *Ricci flow* informally as the process of *stretching* the metric g in directions of negative Ricci curvature, and *contracting* the metric in directions of positive Ricci curvature. The stronger the curvature, the faster the stretching or contracting of the metric. The concepts of stretching and contracting will not be defined formally here, but they increase or decrease the distance between points along these directions. By changing the notion of distance, one also affects the notions of angle and volume (though it turns out that Ricci flow in two dimensions is *conformal*, which means that the notion of angle remains unaffected by the flow; this fact is closely related to the previously mentioned fact that in two dimensions the Ricci curvature is the same in all directions). Ricci flow can be described succinctly and precisely by the equation

$$\frac{d}{dt}g = -2\text{Ric},$$

although we will not define here exactly what it means to differentiate the metric g with respect to the time variable t , or what it means for that derivative to equal the Ricci curvature multiplied by -2 .

In principle, one could perform Ricci flow on a manifold for as long a period of time as one wished. In practice, however, it is possible (especially in the presence of positive curvature) for the Ricci flow to cause a manifold to develop *singularities*: points where it ceases to look like a manifold, and where the geometry may stop resembling Euclidean geometry even at very small scales. For example, if one starts with a perfectly round sphere and performs Ricci flow, what happens is that the sphere contracts at a steady rate until it becomes a point, which is no longer a two-dimensional manifold. In three dimensions, more complicated singularities are possible: for instance, one can have a *neck pinch*, in which a cylinder-like “neck” of the manifold shrinks under Ricci flow, until at one or more places along the neck, the cylinder has tapered down to a point. The types of possible singularity formations for three-dimensional Ricci flow were only classified com-

pletely in a recent and very important paper of Grigori Perelman.

Some years ago, Richard Hamilton made the fundamental observation that Ricci flow is an excellent tool for simplifying the structure of a manifold: generally speaking, it compresses all the positive-curvature parts of the manifold into nothingness, while expanding the negative-curvature parts of the manifold until they become very homogeneous, in the sense that the manifold begins to look much the same no matter which vantage point one selects inside it. Indeed, the flow seems to separate the manifold into extremely symmetric components. For instance, in two dimensions the Ricci flow always ends up endowing the manifold with a metric of constant curvature, which could be positive (as in the sphere), zero (as in the cylinder), or negative (as in *hyperbolic space*); the fact that such a constant-curvature metric can always be found is known as the UNIFORMIZATION THEOREM [V.37] and is of fundamental importance in the theory of surfaces. In higher dimensions, the Ricci flow can develop singularities before perfect symmetry is attained, but it turns out that it is possible to perform “surgeries” (see DIFFERENTIAL TOPOLOGY [IV.9 §§2.3, 2.4]) on the singularities that develop this way, so that the manifold becomes smooth again and one can restart the Ricci flow process. (The surgery may, however, change the topology of the manifold: for instance, it can convert a connected manifold into two disconnected pieces.) In three dimensions it has recently been shown by Perelman that Ricci flow, when augmented by surgery to remove the singularities, does indeed convert an arbitrary manifold (obeying some mild assumptions) into a finite union of some very symmetric and explicitly describable pieces; the precise statement of this conclusion was known as the *geometrization conjecture* of Thurston. One consequence of this conjecture, which is now a rigorous theorem proved by Perelman, is the POINCARÉ CONJECTURE [V.28]: any compact three-dimensional manifold that is *simply connected* (meaning that any closed loop on the manifold can be contracted smoothly to a point without ever leaving the manifold) can in fact be smoothly deformed into a 3-sphere (which is to four-dimensional Euclidean space as the usual two-dimensional sphere is to three-dimensional Euclidean space). The proof of Poincaré’s conjecture is one of the most impressive recent achievements of modern mathematics.

III.81 Riemann Surfaces

Alan F. Beardon

Let D be a *region* (that is, a connected open set) in the complex plane. If f is a complex-valued function defined on D , then we can define its derivative just as we would for real-valued functions defined on subsets of \mathbb{R} : the derivative of f at w is the limit as z tends to w of the “difference quotient” $(f(z) - f(w))/(z - w)$. Of course, this limit does not necessarily exist, but if it exists for every w in D , then f is said to be *analytic*, or *holomorphic*, on D . Analytic functions have amazing properties; for example, if a function is analytic in a region, then it automatically has a Taylor-series expansion at each point of the region, from which one can deduce that it is infinitely differentiable. This is in stark contrast to the theory of real functions of a real variable, where, for example, a function may be once differentiable but not twice differentiable at some point x , yet three-times differentiable at some other point y . *Complex analysis* is the study of analytic functions. Perhaps more than any other mathematical topic, it is both immensely useful in a practical sense and profound and beautiful in a theoretical sense. (See also SOME FUNDAMENTAL MATHEMATICAL DEFINITIONS [I.3 §5.6].)

In general, group theorists do not distinguish between isomorphic groups, and topologists do not distinguish between homeomorphic spaces. Similarly, complex analysts do not distinguish between two regions D and D' if there is an analytic bijection between D and D' . When this is the case, we say that D and D' are *conformally equivalent*. Conformal equivalence is, as its name suggests, an EQUIVALENCE RELATION [I.2 §2.3]: the proof depends on the surprising fact that if f is an analytic bijection from D to D' , then its inverse $f^{-1} : D' \rightarrow D$ is also analytic. Again, this contrasts with real analysis. If D and D' are conformally equivalent, then “interesting” properties of analytic functions on D are transferred automatically to corresponding properties of analytic functions defined on D' . Indeed, this statement can almost be taken as a definition of “interesting” properties (although admittedly this conflicts with the numerical side of complex analysis, because purely numerical statements do not usually transfer under such maps). Naturally, we would like to know which properties of analytic functions are “interesting” in this sense. One such property is that (except at certain isolated points) the angle between two intersecting curves in D is preserved

under an analytic map: this is the origin of the term “conformal.” It is less well-known that if a bijection (which is not assumed to be differentiable) preserves the angles between curves (that is, both their magnitude and whether they are measured clockwise or counterclockwise), then it is analytic. Thus, loosely speaking, the preservation of angles implies the existence of a Taylor series!

The impact of complex analysis on other topics is so great that it is natural to try to find the most general type of surface on which we can study analytic functions. This leads to the definition of a *Riemann surface* (after BERNHARD RIEMANN [VI.49], who introduced the idea in his doctoral dissertation). In order to put a coordinate system on a surface S we try to map S bijectively onto a plane region D ; if we succeed, then we can transfer the coordinates from D to S . For many surfaces (for example, a sphere) it is not possible to find such a map, and we have to be satisfied with *local coordinates*. This means that at each point w of S , we map a neighborhood N of w onto a plane region, and so obtain coordinates that are restricted to N . As there are usually infinitely many ways to do this, we are forced to consider the class of *transition maps*; that is, the maps from one coordinate system at w to another. The surface is a *Riemann surface* precisely when each such transition map is an analytic bijection. This definition resembles that of a two-dimensional MANIFOLD [I.3 §6.9], but the requirement that the transition maps should be analytic is much stronger, so by no means every 2-manifold is a Riemann surface.

It is not difficult to construct Riemann surfaces. Consider, for example, a sphere S resting on a horizontal table. If we imagine a light source at the highest point P of the sphere, then each point of S except P casts a “shadow” on the table: since the table has a simple coordinate system, we can use these “shadows” to define a coordinate system on all of S except the point P . Similarly, a light source at the point Q of tangency with the table casts a shadow onto the (horizontal) tangent plane at P , and this gives a coordinate system valid throughout S except at Q . It can be shown that if the second coordinate system is composed with a reflection, then the sphere does have the structure of a Riemann surface. This is an extremely important example, because it allows one to handle questions involving infinity in a satisfactory way; it is known as the *Riemann sphere*.

For another example, consider a cube C , and (for simplicity only) remove the eight vertices. Given a face F of

C (without its bounding edges), we can find a Euclidean rigid motion that maps F into \mathbb{C} , so we can easily define a coordinate system on F . If w is an interior point of an edge E of C , we can “open” the two faces that meet at E to make a planar region that contains E , and then map this region into \mathbb{C} by a Euclidean rigid motion. In this way we see that C (less its vertices) is a Riemann surface. The problem with the vertices can be solved by technical means, and this method can then be generalized to show that any polyhedron (even one with holes, such as a “square” torus) is a Riemann surface. These are known as *compact surfaces*. It is a deep but fascinating classical result that each such surface corresponds bijectively to an irreducible polynomial $P(z, w)$ in two complex variables. To give an idea of how the correspondence works, let us consider an equation such as $w^3 + wz + z^2 = 0$. For each z this can be solved to give three values of w , say w_1 , w_2 , and w_3 ; as we allow z to vary in \mathbb{C} , the values w_j vary, and as they do so they create a Riemann surface W , which can be shown to be connected. This surface can be thought of as lying “above” \mathbb{C} , and for all but a finite set of z in \mathbb{C} there are exactly three points on W that are “above” z .

As we have mentioned, Riemann surfaces are important because they are the most general surfaces on which one can study analytic functions, with all of their remarkable properties. It is easy to define what we mean by an analytic function f on a Riemann surface R . Given a coordinate system on part of R , we can think of f as a function of the coordinates, and we then regard f as analytic if and only if it depends analytically on the coordinates. Because the transition maps are analytic, f will be analytic with respect to one coordinate system if and only if it is analytic with respect to all the other coordinate systems defined at the point in question.

This simple property—that if something holds in one coordinate system, then it holds in all of them—is one of the crucial features of the theory. For example, suppose that we have two curves crossing on an (abstract) Riemann surface. If we transfer the two curves to plane regions using different local coordinate systems at the crossing point, and then measure the angle of intersection in each case, we must get the same result (since the transition from one coordinate system to another preserves angles). It follows that the angle between intersecting curves on an abstract Riemann surface is a well-defined concept.

It turns out that analysis on Riemann surfaces goes beyond analytic functions. *Harmonic functions* (solutions of LAPLACE’S EQUATION [I.3 §5.4]) are intimately

connected to analytic functions, since the real part of an analytic function is harmonic and any harmonic function is (locally) the real part of an analytic function. Thus, on a Riemann surface, complex analysis merges almost imperceptibly with potential theory (which is the study of harmonic functions).

Perhaps the most profound theorem of all about Riemann surfaces is the UNIFORMIZATION THEOREM [V.37]. Roughly speaking, this says that every Riemann surface is obtained from either Euclidean, spherical, or hyperbolic geometry (see SOME FUNDAMENTAL MATHEMATICAL DEFINITIONS [I.3 §§6.2, 6.5, 6.6]) by taking a polygon in that geometry and gluing its sides together, in the same way that one obtains a torus by gluing opposite sides of a rectangle together. (See also FUCHSIAN GROUPS [III.28].) Remarkably, only very few Riemann surfaces come from the Euclidean or spherical geometries; essentially, every Riemann surface can be constructed in this way from (and only from) the hyperbolic plane. This means that virtually every region in the complex plane comes equipped with a natural and intrinsic geometry whose character is hyperbolic and *not*, as one might expect, Euclidean. The Euclidean character of a generic plane region comes from its embedding in \mathbb{C} , and not from its own intrinsic hyperbolic geometry.

III.82 The Riemann Zeta Function

The *Riemann zeta function* ζ is a function defined on the complex numbers that encapsulates in a remarkable way many of the most important properties about the distribution of prime numbers. If s is a complex number with real part greater than 1, then $\zeta(s)$ is defined to be $\sum_{n=1}^{\infty} n^{-s}$. The condition that $\operatorname{Re}(s) > 1$ is needed to ensure that this series converges. However, because the resulting function is HOLOMORPHIC [I.3 §5.6], it is possible to extend the definition by means of analytic continuation. The result is a function that is defined everywhere on the complex plane (though it takes the value ∞ at 1).

A first clue to why this function is related to the distribution of primes is *Euler's product formula*:

$$\zeta(s) = \prod_p (1 - p^{-s})^{-1}.$$

Here, the product on the right-hand side is over all primes. The formula can be proved by writing $(1 - p^{-s})^{-1}$ as $1 + p^{-s} + p^{-2s} + \dots$, expanding out the product, and using THE FUNDAMENTAL THEOREM OF ARITHMETIC [V.16]. Deeper connections were discovered by

RIEMANN [VI.49], who formulated the famous RIEMANN HYPOTHESIS [IV.4 §3].

The Riemann zeta function is just one of a family of functions that encode important number-theoretic information. For example, *Dirichlet L-functions* are closely related to the distribution of primes in arithmetic progressions. For more details about these and about the Riemann zeta function itself, see ANALYTIC NUMBER THEORY [IV.4]. Some more sophisticated zeta functions are described in THE WEIL CONJECTURES [V.38]. See also L-FUNCTIONS [III.49].

III.83 Rings, Ideals, and Modules

1 Rings

A ring, like a GROUP [I.3 §2.1] or a FIELD [I.3 §2.2], is an algebraic structure that satisfies certain axioms. To remember the axioms for both rings and fields at the same time, it is helpful to think of two simple examples: with the two operations of addition and multiplication, the set \mathbb{Z} of all integers forms a ring and the set \mathbb{Q} of all rational numbers forms a field. In general, a ring is a set R with two BINARY OPERATIONS [I.2 §2.4], denoted by “+” and “×”, which satisfies all the field axioms apart from the one that says that nonzero elements have multiplicative inverses.

Although the integers are the prototypical example of a ring, the notion arose historically as an abstraction from several sources, one of which was polynomials. Like integers, polynomials (with real coefficients, say) can be added and multiplied, and these operations have all the properties one might expect, such as the fact that multiplication is distributive over addition, so the space of such polynomials forms a ring. Other examples include the integers modulo n (for any positive integer n), the rationals (or indeed any other field), and the set $\mathbb{Z}[i]$ of all complex numbers $a + bi$ such that a and b are integers.

Sometimes the assumptions that multiplication is commutative and has an identity element are dropped. This leads to a more complicated theory, but it does encompass important examples such as the set of all $n \times n$ matrices (with elements in a given field, or even just a ring).

As with other algebraic structures, there are several ways of forming new rings from old ones: for instance, we can take subrings and direct products of two rings. Slightly less obviously, we can start with a ring R and

PUP: please note that this article was called 'The Zeta Function' and appeared as the last article in the part, but this is definitely a better name for it and therefore a better position.

form the ring of all polynomials with coefficients in R . We can also take QUOTIENTS [I.3 §3.3], but in order to discuss these we must introduce the notion of an ideal.

2 Ideals

A typical quotient construction for an algebraic structure A will identify some substructure B and regard two elements of A as “equivalent” if they “differ by an element of B .” If A is a group or a VECTOR SPACE [I.3 §2.3], then B will be a subgroup or a subspace. However, the situation for rings is slightly different.

We can see why if we think about quotients in another way: as images of HOMOMORPHISMS [I.3 §4.1]. The substructures that we like to quotient by are the kernels of these homomorphisms, so we should ask ourselves what the kernel of a ring homomorphism (that is, the set of elements that map to 0) will be like.

If $\phi : R \rightarrow R'$ is a homomorphism between two rings, and $\phi(a) = \phi(b) = 0$, then $\phi(a + b) = 0$. Also, if r is any element of R , then $\phi(ra) = \phi(r)\phi(a) = 0$. Thus, the kernel of a homomorphism is closed under addition, and also under multiplication by any element of the ring. These two properties define the notion of an *ideal*. For example, the set of all even integers is an ideal in \mathbb{Z} . In interesting cases, ideals are not subrings, since if an ideal contains 1 then it must contain r for every r in the ring. (An example that makes the difference very clear is the subset of the ring of all polynomials that consists of all constant polynomials. The constants form a subring, but they certainly do not form an ideal.)

It is not hard to show that for any ideal I in a ring R there is a homomorphism that has I as its kernel, namely the quotient map from R to the quotient R/I . Here R/I is a construction that as usual we think of as “ R , but with two elements considered the same if they differ by an element of I .”

Quotients of rings are extremely useful in ALGEBRAIC NUMBER THEORY [IV.3] because they allow us to rephrase questions about algebraic numbers as questions about polynomials. To get an idea of how this is done, consider the ring $\mathbb{Z}[X]$ of all polynomials with integer coefficients, and the ideal that consists of all multiples (by integer polynomials) of the polynomial $X^2 + 1$. In the quotient of $\mathbb{Z}[X]$ by this ideal, we regard two polynomials as the same if they differ by a multiple of $X^2 + 1$. In particular, X^2 is the same as -1 . In other words, in this quotient ring we have a square root of -1 , and in fact the quotient ring is isomorphic to the ring $\mathbb{Z}[i]$ that we met earlier.

One of the things we like to do to integers is factorize them, and we can try to do the same in rings as well. However, it turns out that, while it is usually possible to factorize an element of a ring into “irreducible” ones that cannot be factorized further (like the primes in \mathbb{Z}), in many cases the factorization is not unique. At first, this might be rather unexpected, and indeed it was a stumbling block for many early workers (in the eighteenth and nineteenth centuries). Here is an example: in the ring $\mathbb{Z}[\sqrt{-3}]$, which consists of all complex numbers $a + b\sqrt{-3}$, where a and b are integers, the number 4 may be factorized as 2×2 and also as $(1 + \sqrt{-3}) \times (1 - \sqrt{-3})$.

3 Modules

Modules are to rings as vector spaces are to fields. In other words, they are algebraic structures where the basic operations are addition and scalar multiplication, but now the scalars are allowed to come from a ring rather than a field. For an example of a module over a ring that is not a field, take any Abelian group G . This can be turned into a module over \mathbb{Z} , with addition given by the group operation and scalar multiplication defined in the obvious way: for instance, $3g$ means $g + g + g$, and $-2g$ means the inverse of $g + g$.

The simplicity of this definition masks the fact that the structure of modules is in general far more subtle than that of vector spaces. For example, we can define a *basis* of a module to be a linearly independent set of elements that spans the module. However, many useful facts about bases in vector spaces do not hold for modules. For instance, in \mathbb{Z} , which we may consider as a module over itself, the set $\{2, 3\}$ spans the module but does not contain a basis, and similarly the set $\{2\}$ is linearly independent but cannot be extended to a basis. In fact, modules may be very far from having a basis: for example, if we consider the integers modulo n as a module over \mathbb{Z} , then even a single element x fails to be linearly independent, since $nx = 0$.

The following example of a module is an important one. Let V be a complex vector space and let α be a linear map from V to V . This can be made into a module over the ring $\mathbb{C}[X]$: if $v \in V$ and P is a complex polynomial, then Pv is defined to be $P(\alpha)v$. (For instance, if P is the polynomial $x^2 + 1$, then $Pv = \alpha^2 v + v$.) Applying general structural results about modules to this example, one obtains a proof of the JORDAN NORMAL FORM THEOREM [III.45].

III.84 Schemes

Jordan S. Ellenberg

One frequently finds in the history of mathematics that a definition thought to be completely general was in fact too restrictive to treat certain problems of interest. The notion of “number,” for instance, has been expanded again and again—most notably to incorporate irrationalities and complex numbers, the former arising from problems in geometry and the latter needed in order to describe solutions to arbitrary algebraic equations. In a similar way, algebraic geometry, which was once understood as the study of *algebraic varieties*, or solution sets of algebraic equations in some finite-dimensional space, has grown to encompass more general objects known as “schemes.” As a very meager example one might consider the two equations $x + y = 0$ and $(x + y)^2 = 0$. The two equations have the same set of solutions in the plane, so they describe the same variety; but the schemes attached to the two objects are completely different. The reformulation of algebraic geometry in the language of schemes was a tremendous project spearheaded by Alexander Grothendieck in the 1960s. As the above example suggests, the scheme-theoretic viewpoint tends to emphasize the algebraic aspects of the subject (equations) rather than the traditionally geometric ones (solution sets of equations). This viewpoint has made a reality of the long-hoped-for unification of ALGEBRAIC NUMBER THEORY [IV.3] and algebraic geometry, and, indeed, much recent progress in number theory would have been impossible without the geometric insight supplied by the theory of schemes.

Even schemes are not enough to handle all the problems of current interest, and still more general notions (stacks, “noncommutative varieties,” derived categories of sheaves, etc.) are brought to bear when necessary. These can appear exotic, but to our successors they will no doubt be second nature, just as schemes are to us. For more on algebraic geometry in general, see ALGEBRAIC GEOMETRY [IV.7]. Schemes are discussed at greater length in ARITHMETIC GEOMETRY [IV.6].

III.85 The Schrödinger Equation

Terence Tao

In mathematical physics, the Schrödinger equation (and the closely related Heisenberg equation) are the

most fundamental equations in nonrelativistic quantum mechanics, playing the same role as Hamilton’s laws of motion (and the closely related Poisson equation) in nonrelativistic classical mechanics. (In relativistic quantum mechanics, the equations of quantum field theory take over the role of Heisenberg’s equation, while Schrödinger’s equation does not have a natural direct analogue.) In pure mathematics, the Schrödinger equation, together with its variants, is one of the basic equations studied in the field of PARTIAL DIFFERENTIAL EQUATIONS [IV.16], and has applications to geometry, to spectral and scattering theory, and to integrable systems.

The Schrödinger equation can be used to describe the quantum dynamics of many-particle systems under the influence of a variety of forces, but for simplicity let us consider just a single particle, of mass $m > 0$, moving about in n -dimensional space \mathbb{R}^n subject to the influence of a potential, which we shall take to be a function $V : \mathbb{R}^n \rightarrow \mathbb{R}$. To avoid technicalities we shall assume that all the functions we discuss are smooth.

In classical mechanics, this particle would have a specific position $q(t) \in \mathbb{R}^n$ and a specific momentum $p(t) \in \mathbb{R}^n$ for each time t . (Eventually we shall observe the familiar law $p(t) = mv(t)$, where $v(t) = q'(t)$ is the velocity of the particle.) Thus the state of this system at any given time t is described by the element $(q(t), p(t))$ of the space $\mathbb{R}^n \times \mathbb{R}^n$, which is known as *phase space*. The *energy* of this state is described by the HAMILTONIAN FUNCTION [III.35] $H : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ on phase space, defined in this case by

$$H(q, p) = \frac{|p|^2}{2m} + V(q).$$

(Physically, the quantity $|p|^2/2m = \frac{1}{2}m|v|^2$ represents kinetic energy, while $V(q)$ represents potential energy.) The system then evolves according to *Hamilton’s equations of motion*:

$$q'(t) = \frac{\partial H}{\partial p}, \quad p'(t) = -\frac{\partial H}{\partial q}, \quad (1)$$

where we keep in mind that p and q are vectors, so that these derivatives are GRADIENTS [I.3 §5.3]. Hamilton’s equations of motion are valid for any classical system, but in our specific case of a particle in a “potential well,” they become

$$q'(t) = \frac{1}{m}p(t), \quad p'(t) = -\nabla V(q). \quad (2)$$

The first equation is asserting that $p = mv$, while the second equation is basically Newton’s second law of motion.

From (1) we can easily derive *Poisson's equation of motion*

$$\frac{d}{dt}A(q(t), p(t)) = \{H, A\}(q(t), p(t)) \quad (3)$$

for any *classical observable* $A : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$, where

$$\{H, A\} = \frac{\partial H}{\partial p} \frac{\partial A}{\partial q} - \frac{\partial A}{\partial p} \frac{\partial H}{\partial q}$$

is the *Poisson bracket* of H and A . Setting $A = H$, we have in particular the *conservation-of-energy law*:

$$H(q(t), p(t)) = E \quad (4)$$

for all $t \in \mathbb{R}$ and some quantity E independent of t .

Now we analyze the quantum mechanical analogue of the above classical system. We need a small¹ parameter $\hbar > 0$, known as *Planck's constant*. The state of the particle at a time t is no longer described by a single point $(q(t), p(t))$ in phase space, but is instead described by a *wave function*, which is a complex-valued function of position that evolves over time: that is, for each t we have a function $\psi(t)$ from \mathbb{R}^n to \mathbb{C} . It is required to obey the normalization condition $\langle \psi(t), \psi(t) \rangle = 1$, where $\langle \cdot, \cdot \rangle$ denotes the inner product

$$\langle \phi, \psi \rangle = \int_{\mathbb{R}^n} \phi(q) \overline{\psi(q)} dq.$$

Unlike a classical particle, a wave function $\psi(t)$ does not necessarily have a specific position $q(t)$. However, it does have an *average position* $\langle q(t) \rangle$, defined as

$$\langle q(t) \rangle = \langle Q\psi(t), \psi(t) \rangle = \int_{\mathbb{R}^n} q |\psi(t, q)|^2 dq.$$

Here, we have written $\psi(t, q)$ for the value of $\psi(t)$ at the point q , and Q is the *position operator*, defined by $(Q\psi)(t, q) = q\psi(t, q)$: that is, Q is the operator that multiplies pointwise by q . Similarly, while ψ does not have a specific momentum $p(t)$, it does have an *average momentum* $\langle p(t) \rangle$, defined as

$$\langle p(t) \rangle = \langle P\psi(t), \psi(t) \rangle = \frac{\hbar}{i} \int_{\mathbb{R}^n} (\nabla_q \psi(t, q)) \overline{\psi(t, q)} dq,$$

where the *momentum operator* P is defined by *Planck's law*

$$P\psi(t, q) = \frac{\hbar}{i} \nabla_q \psi(t, q).$$

Note that the vector $\langle p(t) \rangle$ is real-valued because all the components of P are SELF-ADJOINT [III.52 §3.2]. More generally, given any *quantum observable*, by which we mean a self-adjoint OPERATOR [III.52] A acting on the space $L^2(\mathbb{R}^n)$ of complex-valued square integrable

functions, we can define the *average value* $\langle A(t) \rangle$ of A at time t by the formula

$$\langle A(t) \rangle = \langle A\psi(t), \psi(t) \rangle.$$

The analogue of Hamilton's equations of motion (1) is now the *time-dependent Schrödinger equation*:

$$i\hbar \frac{\partial \psi}{\partial t} = H\psi, \quad (5)$$

where H is now a quantum observable rather than a classical one. More precisely,

$$H = \frac{|P|^2}{2m} + V(Q).$$

In other words, we have

$$\begin{aligned} i\hbar \frac{\partial \psi}{\partial t}(t, q) &= H\psi(t, q) \\ &= -\frac{\hbar^2}{2m} \Delta_q \psi(t, q) + V(q)\psi(t, q), \end{aligned}$$

where

$$\Delta_q \psi = \sum_{j=1}^n \frac{\partial^2 \psi}{\partial q_j^2}$$

is the *Laplacian* of ψ . The analogue of Poisson's equation of motion (3) is the *Heisenberg equation*

$$\frac{d}{dt} \langle A(t) \rangle = \left\langle \frac{i}{\hbar} [H(t), A(t)] \right\rangle \quad (6)$$

for any observable A , where $[A, B] = AB - BA$ is the *commutator* or *Lie bracket* of A and B . (The quantity $(i/\hbar)[A, B]$ is occasionally referred to as the *quantum Poisson bracket* of A and B .)

If the quantum state ψ oscillates in time according to the formula $\psi(t, q) = e^{(E/i\hbar)t} \psi(0, q)$ for some real number E (known as the *energy level* or *eigenvalue*), then one has the *time-independent Schrödinger equation*:

$$H\psi(t) = E\psi(t) \quad \text{for all times } t \quad (7)$$

(compare this with (4)). More generally, the important subject of *spectral theory* provides many links between the time-dependent equation (5) and the time-independent equation (7).

There are several strong analogies between the equations of classical mechanics and those of quantum mechanics. For instance, from (6) one has the equations

$$\frac{d}{dt} \langle q(t) \rangle = \frac{1}{m} \langle p(t) \rangle, \quad \frac{d}{dt} \langle p(t) \rangle = -\langle \nabla_q V(q)(t) \rangle,$$

which should be compared with (2). Also, given any classical solution $t \mapsto (q(t), p(t))$ to Hamilton's equation of motion, one can construct a corresponding

1. In many applications it is convenient to normalize \hbar (and m) to equal 1.

family of *approximate* solutions $\psi(t)$ to Schrödinger's equation, for instance by the formula²

$$\psi(t, q) = e^{(i/\hbar)L(t)} e^{(i/\hbar)p(t) \cdot (q - q(t))} \varphi(q - q(t)),$$

where

$$L(t) = \int_0^t \frac{p(s)^2}{2m} - V(q(s)) \, ds$$

is the *classical action* and φ is any slowly varying function that is normalized in the sense that $\int_{\mathbb{R}^n} |\varphi(q)|^2 \, dq = 1$. One can verify that ψ solves (5) except for some errors that are small when \hbar is small. In physics, this fact is an example of the *correspondence principle*, which asserts that classical mechanics can be used to approximate quantum mechanics accurately if Planck's constant is small and one is working at *macroscopic* scales (which is what allows us to use slowly varying functions φ). In mathematics (and more precisely in the fields of *microlocal analysis* and *semi-classical analysis*), there are a number of formalizations of this principle that allow us to use knowledge about the behavior of Hamilton's equations of motion in order to analyze the Schrödinger equation. For example, if the classical equations of motion have periodic solutions, then the Schrödinger equation often has nearly periodic solutions, whereas if the classical equations have very chaotic solutions, then the Schrödinger equation typically does as well (this phenomenon is known as *quantum chaos* or *quantum ergodicity*).

There are many aspects of the Schrödinger equation that are of interest. We mention just one of them here for illustration, namely that of *scattering theory*. If the potential function V decays sufficiently quickly at infinity, and $k \in \mathbb{R}^n$ is a nonzero frequency vector, then, setting the energy level as $E = \hbar^2 |k|^2 / 2m$, the *time-dependent* Schrödinger equation $H\psi = E\psi$ admits solutions $\psi(q)$ that behave asymptotically (as $|q| \rightarrow \infty$) as

$$\psi(q) \approx e^{ik \cdot q} + f\left(\frac{q}{|q|}, k\right) \frac{e^{i|k||q|}}{r^{(n-1)/2}}$$

for some canonical function $f : S^{n-1} \times \mathbb{R}^n \rightarrow \mathbb{C}$, which is known as the *scattering amplitude function*. This scattering amplitude depends (in a nonlinear fashion) on the potential V , and the map from V to f is known as

2. Intuitively, this function $\psi(t, q)$ is localized in position near $q(t)$ and localized in momentum near $p(t)$, and is thus localized near $(q(t), p(t))$ in phase space. Such a localized function, exhibiting such "particle-like" behavior as having a reasonably well-defined position and velocity, is sometimes known as a "wave packet." A typical solution of the Schrödinger equation does not behave like a wave packet, but can be decomposed as a *superposition* or *linear combination* of wave packets; such decompositions are a useful tool in analyzing general solutions of such equations.

the *scattering transform*. The scattering transform can be viewed as a nonlinear variant of THE FOURIER TRANSFORM [III.27]; it is connected to many areas of partial differential equations, such as the theory of *integrable systems*.

There are many generalizations and variants of the Schrödinger equation; one can generalize to many-particle systems, or add other forces such as magnetic fields or even nonlinear terms. One can also couple this equation to other physical equations such as MAXWELL'S EQUATIONS [IV.17 §1.1] of electromagnetism, or replace the domain \mathbb{R}^n by another space such as a torus, a discrete lattice, or a manifold. Alternatively, one could place some impenetrable obstacles in the domain (thus effectively removing those regions of space from the domain). The study of all of these variants leads to a vast and diverse field in both pure mathematics and in mathematical physics.

III.86 The Simplex Algorithm

Richard Weber

1 Linear Programming

The simplex algorithm is the preeminent tool for solving some of the most important mathematical problems arising in business, science, and technology. In these problems, which are called linear programs, we are to maximize (or minimize) a linear function subject to linear constraints. An example is the diet problem posed by the U.S. Air Force in 1947: find quantities of seventy-seven differently priced foodstuffs (cheese, spinach, etc.) to satisfy a man's minimum daily requirements for nine nutrients (protein, iron, etc.) at least cost. Further applications occur in choosing the elements of an investment portfolio, rostering an airline's crew, and finding optimal strategies in two-person games. The study of linear programming has inspired many of the central ideas of optimization theory, such as DUALITY [III.19], the importance of convexity, and COMPUTATIONAL COMPLEXITY [IV.21].

The input data of a linear program (LP) consists of two vectors $b \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$, and an $m \times n$ matrix $A = (a_{ij})$. The problem is to find values for n non-negative decision variables, x_1, \dots, x_n , to maximize the *objective function* $c_1 x_1 + \dots + c_n x_n$, subject to m constraints, $a_{i1} x_1 + \dots + a_{in} x_n \leq b_i$, $i = 1, \dots, m$. In the diet problem, $n = 77$ and $m = 9$. In the following simple example (not a diet problem), $n = 2$ and $m = 3$. In serious real-life problems, n and m can be greater than

PUP: this is indeed OK as written.

PUP: Tim has checked this against usage earlier in article and confirms it's OK.

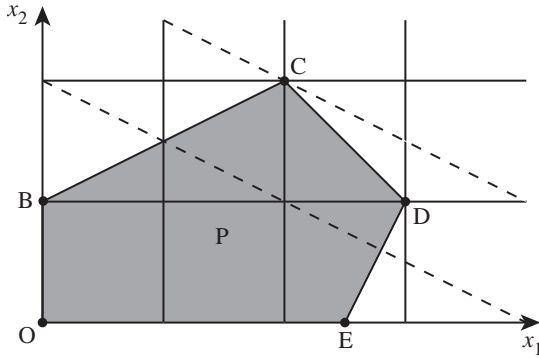


Figure 1 Feasible region “P” of an LP.

100 000.

$$\begin{aligned}
 &\text{Maximize} && x_1 + 2x_2 \\
 &\text{subject to} && -x_1 + 2x_2 \leq 2, \\
 & && x_1 + x_2 \leq 4, \\
 & && 2x_1 - x_2 \leq 5, \\
 & && x_1, x_2 \geq 0.
 \end{aligned}$$

The constraints define a feasible region for (x_1, x_2) , a convex polygon that is depicted by the shaded region “P” in figure 1. The two dotted lines mark those x where the value of the objective function value is 4 and where it is 6. Clearly, it is maximized at point C.

The general story is similar to that of the example. If the feasible region $P = \{x : Ax \leq b, x \geq 0\}$ is nonempty, then it is a convex polytope in \mathbb{R}^n , and an optimal solution can be found at one of its vertices. It is helpful to introduce “slack variables” $x_3, x_4, x_5 \geq 0$ to take up the slack on the left of the inequality constraints. We can write

$$\begin{aligned}
 -x_1 + 2x_2 + x_3 &= 2, \\
 x_1 + x_2 + x_4 &= 4, \\
 2x_1 - x_2 + x_5 &= 5.
 \end{aligned}$$

We now have three equations in five variables, so we can set any two of the variables x_1, \dots, x_5 equal to 0, and solve the equations for the other three variables (or solve a perturbation of them if they happen not to be independent). There are ten ways to choose two variables from five. Not all of the ten corresponding solutions satisfy $x_1, x_2, x_3, x_4, x_5 \geq 0$, but five of them do. These are called *basic feasible solutions* (BFSs), and correspond to the vertices of P marked O, B, C, D, E.

2 How the Algorithm Works

George Dantzig invented the simplex algorithm in 1947 as a means of solving the Air Force’s diet problem mentioned at the start. The word “program” was not yet used to mean computer code, but was a military term for a logistic plan or schedule. The fundamental fact on which the algorithm relies is that if an LP has a bounded optimal solution, then the optimum value is attained at a BFS, i.e., at a vertex (or so-called “extreme point”) of the polytope of feasible points, P. Another name for the feasible polytope is “simplex,” which is where the algorithm gets its name. It works as follows.

Step 0. Pick a BFS.

Step 1. Test whether this BFS is optimal.

If so, stop. If not, go to step 2.

Step 2. Find a better BFS.

Repeat from step 1.

Since there are only finitely many BFSs (i.e., vertices of P), the algorithm must stop.

Now that we have an overview, let us look at the details. Suppose that at step 0 we pick the BFS of $x = (x_1, x_2, x_3, x_4, x_5) = (0, 0, 2, 4, 5)$, corresponding to vertex O. At step 1 we wish to know if the objective function can be increased if x_1 or x_2 is increased from 0. So we write x_3, x_4, x_5 , and the objective function $c^T x$ in terms of x_1 and x_2 , and display this as dictionary 1.

Dictionary 1	
x_3	$= 2 + x_1 - 2x_2,$
x_4	$= 4 - x_1 - x_2,$
x_5	$= 5 - 2x_1 + x_2,$
$c^T x$	$= x_1 + 2x_2.$

The last equation in the dictionary shows that we can increase the value of $c^T x$ by increasing either x_1 or x_2 from 0. Suppose that we increase x_2 . The first and second equations show that x_3 and x_4 must decrease, and we cannot increase x_2 beyond 1, at which point $x_3 = 0$ and $x_4 = 3, x_5 = 6$. Increasing x_2 as much as possible, we complete step 2 and arrive at a new BFS of $x = (0, 1, 0, 3, 6)$, which is vertex B. Now we are ready for step 1 again, and so we write x_2, x_4, x_5 , and $c^T x$ in terms of the variables that are now zero, namely x_1, x_3 , to give dictionary 2.

Dictionary 2	Dictionary 3
$x_2 = 1 + \frac{1}{2}x_1 - \frac{1}{2}x_3,$	$x_1 = 2 + \frac{1}{3}x_3 - \frac{2}{3}x_4,$
$x_4 = 3 - \frac{3}{2}x_1 + \frac{1}{2}x_3,$	$x_2 = 2 - \frac{1}{3}x_3 - \frac{1}{3}x_4,$
$x_5 = 6 - \frac{3}{2}x_1 - \frac{1}{2}x_3,$	$x_5 = 3 - x_3 + x_4,$
$c^T x = 2 + 2x_1 - x_3.$	$c^T x = 6 - \frac{1}{3}x_3 - \frac{4}{3}x_4.$

This shows that $c^T x$ can be increased by increasing x_1 from 0, but that x_1 can increase no further than 2 because at that point $x_4 = 0$. This brings us to a new solution $(2, 2, 0, 0, 3)$, which is vertex C. Once more, we are ready for step 1, and so compute dictionary 3, now writing things in terms of x_3 and x_4 , which are 0. The algorithm now stops because, as we require $x_3, x_4 \geq 0$, the bottom line of dictionary 3 proves that $c^T x \leq 6$ for all feasible x .

There is other important information in the final dictionary. If b is changed to $b + \epsilon$, for small $\epsilon^T = (\epsilon_1, \epsilon_2, \epsilon_3)$, then the maximum value of $c^T x$ will change to $6 + \frac{1}{3}\epsilon_1 + \frac{4}{3}\epsilon_2$. The coefficient $\frac{1}{3}$ is called a “shadow price,” because it is what we should be willing to pay per unit increase in b_1 .

3 How the Algorithm Performs

In running the simplex algorithm the serious work comes in computing the dictionaries. To find dictionary 2, we could use the first equation of dictionary 1 to rewrite x_2 in terms of x_1 and x_3 , and then substitute for x_2 in the other equations. Versions of the simplex algorithm have been invented that reduce the computing effort by taking advantage of special structure in the matrix A , such as the fact that most of its entries are zero. The dictionary data is often held in a so-called *tableau* of coefficients.

There are many other practical and theoretical issues. One concerns the selection of the *pivot*, that is, the variable that is to be increased from 0. Starting at O, and depending on which of x_1 and x_2 we choose as the first variable to increase from zero, the path to C can be O, E, D, C or O, B, C. There is no known way to guarantee that the algorithm takes the shortest path.

The question of how many steps the simplex algorithm really needs is related to the famous *Hirsch conjecture*: that for any bounded n -dimensional polytope with m faces, the diameter (defined as the maximum number of edges on the shortest edge-traversing path between any two vertices) is at most $m - n$. If this were true, it would suggest that some version of the simplex algorithm might run in a number of steps that

grows only linearly in the numbers of variables and constraints. However, Klee and Minty (1972) have given an example based on a perturbed n -dimensional cube ($m = 2n$ faces and diameter n), in which if the algorithm selects among possible pivots by choosing the one for which the objective function increases at the greatest rate per unit increase in that variable, then it visits all 2^n vertices before reaching the optimum. Indeed, for most deterministic pivot selection rules, examples are known in which the number of steps grows exponentially in n .

Fortunately, things are usually much better in practical problems than in worst-case examples. Typically, only $O(m)$ steps are needed to solve a problem with m constraints. Moreover, Khachian (1979) proved (by analysis of the so-called *ellipsoid algorithm*) that linear programs can in principle be solved by an algorithm whose running time grows only polynomially in n . Thus linear programming is much easier than “integer linear programming,” in which x_1, \dots, x_n are required to be integers and for which no algorithm with polynomial running time is known.

Karmarkar (1984) pioneered development of “interior” methods for linear programming problems. These move through the interior of the polytope P , rather than among its vertices, and can sometimes solve large LPs more quickly than the simplex algorithm. Modern computer software uses both methods and can easily solve LPs with millions of variables and constraints.

Further Reading

- Dantzig, G. 1963. *Linear Programming and Extensions*. Princeton, NJ: Princeton University Press.
- Karmarkar, N. 1984. A new polynomial-time algorithm for linear programming. *Combinatorica* 4:373–95.
- Khachian, L. G. 1979. A polynomial algorithm in linear programming. *Soviet Mathematics Doklady* 20:191–94.
- Klee, V., and G. Minty. 1972. How good is the simplex algorithm? In *Inequalities III*, edited by O. Shisha, volume 16, pp. 159–75. New York: Academic Press.

Solitons

See LINEAR AND NONLINEAR WAVES AND SOLITONS [III.51]

III.87 Special Functions

T. W. Körner

Suppose that the only functions we have come across are quotients of polynomials and that we are asked to solve the differential equation

$$f'(x) = 1/x \quad (1)$$

for all $x > 0$, subject to the condition $f(1) = 0$.

If we try $f(x) = P(x)/Q(x)$, where P and Q are polynomials with no common factors, then we find that

$$x(Q(x)P'(x) - P(x)Q'(x)) = Q(x)^2.$$

By comparing coefficients we can show that $Q(0) = P(0) = 0$, which shows that, contrary to our assumptions, both $P(x)$ and $Q(x)$ are divisible by x . Thus, we cannot solve equation (1) in terms of known functions. However, THE FUNDAMENTAL THEOREM OF CALCULUS [I.3 §5.5] tells us that equation (1) does indeed have a solution, namely

$$F(x) = \int_1^x \frac{1}{t} dt.$$

Further study shows that the function F has many useful properties. For example, using the substitution $u = t/a$, we find that

$$\begin{aligned} F(ab) &= \int_1^a \frac{1}{t} dt + \int_a^{ab} \frac{1}{t} dt \\ &= \int_1^a \frac{1}{t} dt + \int_1^b \frac{1}{u} du \\ &= F(a) + F(b), \end{aligned}$$

and, using the formula for differentiating an inverse function, we find that F^{-1} is the solution of the differential equation

$$g'(x) = g(x).$$

We therefore give the function a name (the *logarithm*) and add it to our list of standard functions.

At a more advanced level, integration by parts shows that the GAMMA FUNCTION [III.31] (introduced by EULER [VI.19])

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt,$$

defined for all $x > 0$, has the property that

$$\Gamma(x) = (x-1)\Gamma(x-1)$$

for all $x > 1$, and therefore $\Gamma(n) = (n-1)!$ for all integers $n \geq 1$ (since $\Gamma(1) = 1$). As one might expect from its association with factorials, the gamma function turns out to be very useful in number theory and statistics.

In practice, a “special function” is any function that, like the logarithm and the gamma function, has been extensively studied and has turned out to be useful. Some authors use the phrase “special functions” in a

more restricted sense, meaning something like “functions that turn up in the solution of physical problems” or “functions other than those generally provided by a pocket calculator,” but these restrictions do not seem to be very useful.

In spite of this apparent generality, the theory of special functions is linked in the minds of many mathematicians to a collection of particular ideas and methods. Indeed, it is often linked to particular books like Whittaker and Watson’s *A Course of Modern Analysis* (which was first published in 1902 and is still selling well) and Abramowitz and Stegun’s *Handbook of Mathematical Functions*. These connections may simply be accidents of history, but the phrase “special functions” is often associated with other phrases like “equations of mathematical physics,” “beautiful formulas,” and “sheer ingenuity.” We illustrate this and other themes in the particular case of *Legendre polynomials*. (The next paragraph involves more advanced mathematics and glosses over several long calculations, but the reader may simply glance over its contents and resume careful reading thereafter.)

Suppose that we wish to examine the gravitational potential ψ of Earth by looking at solutions of LAPLACE’S EQUATION [I.3 §5.4] $\Delta\psi = 0$. Since Earth is more or less spherical, we use spherical polar coordinates (r, θ, ϕ) and, noting that Earth is symmetric about its axis of rotation, we may suppose that ϕ depends only on r and θ . Under these assumptions, Laplace’s equation takes the form

$$\sin\theta \frac{\partial}{\partial r} \left(r^2 \frac{\partial \psi}{\partial r} \right) + \frac{\partial}{\partial \theta} \left(\sin\theta \frac{\partial \psi}{\partial \theta} \right) = 0. \quad (2)$$

Following the standard technique of separation of variables, we look for solutions of the form $\psi(r, \theta) = R(r)\Theta(\theta)$. After a little calculation, equation (2) yields

$$\frac{1}{R(r)} \frac{d}{dr} (r^2 R'(r)) = - \frac{1}{\sin\theta \Theta(\theta)} \frac{d}{d\theta} (\sin\theta \Theta'(\theta)). \quad (3)$$

Since one side of equation (3) depends on r alone and the other on θ alone, both sides must equal some constant k . The equation

$$\frac{1}{R(r)} \frac{d}{dr} (r^2 R'(r)) = k$$

has the solution $R(r) = r^l$ whenever $l(l+1) = k$. The corresponding equation for Θ is then

$$\frac{1}{\sin\theta \Theta(\theta)} \frac{d}{d\theta} (\sin\theta \Theta'(\theta)) = -l(l+1). \quad (4)$$

We now make the substitution $x = \cos\theta$, $y(x) = \Theta(\theta)$ to convert (4) to *Legendre’s equation*

$$(1-x^2)y''(x) - 2xy'(x) + l(l+1)y(x) = 0. \quad (5)$$

Routine equating of coefficients reveals that, if we seek nontrivial solutions of the form $f(x) = \sum_{j=0}^{\infty} a_j x^j$, then, unless l is an integer, $f(x)$ is unbounded as x approaches 1 (that is, as θ approaches 0), so these solutions are not useful physically. However, if l is a positive integer, then there is a polynomial solution of degree l . (If l is a negative integer, the same polynomials reappear.) In fact, we have the following stronger statement: if l is a positive integer, then there exists a unique polynomial P_l of degree l satisfying Legendre's equation (5) such that $P_l(1) = 1$. We call P_l the l th Legendre polynomial. Returning to our original problem, we see that it has solutions of the form

$$\psi(r, \theta) = \sum_{n=0}^{\infty} A_n \frac{P_n(\cos \theta)}{r^{n+1}}.$$

It is obvious to the physicist, and can be proved by the mathematician, that this is the most general solution if we also demand that $\phi(r, \theta) \rightarrow 0$ as $r \rightarrow \infty$. Notice that if r is large, then only the first few terms will contribute much to the final answer.

There are many different ways of obtaining the Legendre polynomials. The reader is invited to verify that, if we define Q_n inductively by setting $Q_0(x) = 1$ and $Q_1(x) = x$, and using the "three-term recurrence relation"

$$(n+1)Q_{n+1}(x) - (2n+1)xQ_n(x) + nQ_{n-1}(x) = 0,$$

then $Q_n(1) = 1$ and Q_n is a polynomial that satisfies Legendre's equation (5) (with $l = n$), from which it follows that Q_n is the Legendre polynomial of degree n .

If we set $v_n(x) = (x^2 - 1)^n$, then

$$(x^2 - 1)v'_n(x) = 2nxv(x).$$

Differentiating both sides of this equation $n+1$ times using Leibniz's rule, we see that $v_n^{(n)}$ satisfies Legendre's equation (5) with $l = n$. Differentiating $v_n(x) = (x-1)^n(x+1)^n$ n times using Leibniz's rule and noting that all but one of the resulting terms vanish when $x = 1$, we see that $v_n^{(n)}$ is a polynomial with $v_n^{(n)}(1) = 2^n n!$. Putting all this information together, we obtain *Rodriguez's formula*

$$P_n(x) = \frac{1}{2^n n!} v_n^{(n)}(x) = \frac{1}{2^n n!} \frac{d}{dx} (x^2 - 1)^n.$$

Equation (5) is an example of a *Sturm-Liouville equation*. Setting $l = n$ and $y = P_n$ and rewriting slightly, we obtain the equation

$$\frac{d}{dx} ((1-x^2)P'_n(x)) + n(n+1)P_n(x) = 0. \quad (6)$$

If m and n are positive integers, then, using (6) and integrating by parts, we obtain

$$\begin{aligned} & -n(n+1) \int_{-1}^1 P_n(x)P_m(x) dx \\ &= \int_{-1}^1 \left(\frac{d}{dx} ((1-x^2)P'_n(x)) \right) P_m(x) dx \\ &= [(1-x^2)P'_n(x)P_m(x)]_{-1}^1 \\ &\quad + \int_{-1}^1 (1-x^2)P'_n(x)P'_m(x) dx \\ &= \int_{-1}^1 (1-x^2)P'_n(x)P'_m(x) dx. \end{aligned}$$

Thus, by symmetry,

$$\begin{aligned} n(n+1) \int_{-1}^1 P_n(x)P_m(x) dx \\ = m(m+1) \int_{-1}^1 P_n(x)P_m(x) dx, \end{aligned}$$

and, if $m \neq n$,

$$\int_{-1}^1 P_n(x)P_m(x) dx = 0. \quad (7)$$

The "orthogonality relation" given by (7) has important consequences. Since P_r is a polynomial of degree exactly r , we know that any polynomial Q of degree $n-1$ or less can be written

$$Q(x) = \sum_{r=0}^{n-1} a_r P_r(x)$$

and so

$$\int_{-1}^1 P_n(x)Q(x) dx = \sum_{r=0}^{n-1} a_r \int_{-1}^1 P_n(x)P_r(x) dx = 0. \quad (8)$$

Thus, P_n is orthogonal to all polynomials of lower degree.

Suppose that $P_n(x)$ changes sign at the points $\alpha_1, \alpha_2, \dots, \alpha_m$ on the interval $[-1, 1]$. Then, if we write

$$Q(x) = (x - \alpha_1)(x - \alpha_2) \cdots (x - \alpha_m),$$

we know that $P(x)Q(x)$ does not change sign on $[-1, 1]$ and so

$$\int_{-1}^1 P_n(x)Q(x) dx \neq 0.$$

By equation (8) this means that the degree m of Q is at least n and so (since a polynomial of degree n can have at most n zeros) P_n must have exactly n distinct zeros on $[-1, 1]$.

GAUSS [VI.26] made use of these facts to obtain a powerful method of numerical integration. Suppose that x_1, x_2, \dots, x_{n+1} are distinct points on $[-1, 1]$. If we set

$$e_j(x) = \prod_{i \neq j} \frac{x - x_i}{x_i - x_j},$$

then $e_j(x)$ is a polynomial of degree n that takes the value 1 when $x = x_j$ and 0 when $x = x_k$ with $k \neq j$. Thus, if R is any polynomial of degree at most n , the polynomial Q given by

$$Q(x) = R(x_1)e_1(x) + R(x_2)e_2(x) + \cdots + R(x_{n+1})e_{n+1}(x) - R(x)$$

has degree at most n , and $R - Q$ vanishes at the $n + 1$ points x_j . It follows that $R = Q$, so

$$R(x) = R(x_1)e_1(x) + R(x_2)e_2(x) + \cdots + R(x_{n+1})e_{n+1}(x).$$

If we write $a_j = \int_{-1}^1 e_j(x) dx$, then

$$\int_{-1}^1 R(x) dx = a_1 R(x_1) + a_2 R(x_2) + \cdots + a_n R(x_{n+1}).$$

It is natural to hope that the approximation

$$\int_{-1}^1 f(x) dx \approx a_1 f(x_1) + a_2 f(x_2) + \cdots + a_n f(x_{n+1}), \quad (9)$$

which is an exact equality when f is a polynomial of degree n or less, will work well for other well-behaved functions.

Gauss observed that we can make a major improvement by taking the x_j to be the $n + 1$ roots of the $(n + 1)$ st Legendre polynomial. Suppose that P is a polynomial of degree at most $2n + 1$. Then we can write

$$P(x) = Q(x)P_{n+1}(x) + R(x),$$

where Q and R are polynomials of degree at most n and P_{n+1} is the $(n + 1)$ st Legendre polynomial. Now P_{n+1} is orthogonal to polynomials of lower degree (and, in particular, to Q), $P_{n+1}(x_j) = 0$ by the definition of x_j , and the approximation (9) is an equality for R . Thus,

$$\begin{aligned} \int_{-1}^1 P(x) dx &= \int_{-1}^1 P_{n+1}(x)Q(x) dx + \int_{-1}^1 R(x) dx \\ &= 0 + \sum_{j=1}^{n+1} a_j R(x_j) \\ &= \sum_{j=1}^{n+1} a_j (P_{n+1}(x_j)Q(x_j) + R(x_j)) \\ &= \sum_{j=1}^{n+1} a_j P(x_j). \end{aligned}$$

We have shown that the “quadrature formula” (9) is actually exact for all polynomials of degree at most $2n + 1$, provided we choose the x_j to be the numbers suggested by Gauss. Unsurprisingly, this choice gives an extremely good way of estimating integrals numerically. “Gaussian quadrature” is one of the two

main methods used to evaluate integrals on computers today.

We conclude with a brief look at a few other special functions.

Consider de Moivre’s formula

$$\cos n\theta + i \sin n\theta = (\cos \theta + i \sin \theta)^n.$$

Using the binomial expansion, we see that

$$\cos n\theta + i \sin n\theta = \sum_{r=0}^n \binom{n}{r} (i)^r \cos^{n-r} \theta \sin^r \theta,$$

and, taking real parts,

$$\cos n\theta = \sum_{r=0}^n \binom{n}{2r} (-1)^r \cos^{n-2r} \theta \sin^{2r} \theta.$$

Since $\sin^2 \theta = 1 - \cos^2 \theta$, we have

$$\begin{aligned} \cos n\theta &= \sum_{r=0}^n \binom{n}{2r} (-1)^r \cos^{n-2r} (1 - \cos^2 \theta)^r \\ &= T_n(\cos \theta), \end{aligned}$$

where T_n is a polynomial of degree n called the n th *Chebyshev polynomial*. The Chebyshev polynomials play an important role in numerical analysis.

The next collection of functions requires us to calculate with infinite sums. Readers may treat our calculations as plausible or justify them rigorously according to taste. Observe first that

$$h(x) = \sum_{n=-\infty}^{\infty} \frac{1}{(x - n\pi)^2}$$

is well-defined for all real noninteger x . Note also that $h(x + \pi) = h(x)$ and $h(\frac{1}{2}\pi - x) = h(\frac{1}{2}\pi + x)$. Set $f(x) = h(x) - \operatorname{cosec}^2(\pi x)$. By showing that there are constants K_1 and K_2 such that

$$0 < \sum_{n=1}^{\infty} \frac{1}{(x - n\pi)^2} < K_1$$

and

$$0 < \operatorname{cosec}^2 x - \frac{1}{x^2} < K_2$$

for all $0 < x \leq \frac{1}{2}\pi$, we deduce that there is a constant K such that $|f(x)| < K$ for all $0 < x < \pi$. Simple calculations show that

$$f(x) = \frac{1}{4} (f(\frac{1}{2}x) + f(\frac{1}{2}(x + \pi))). \quad (10)$$

A single application of (10) shows that $|f(x)| < \frac{1}{2}K$ for all $0 < x < \pi$, and repeated applications show that $f(x) = 0$. Thus

$$\operatorname{cosec}^2 x = \sum_{n=-\infty}^{\infty} \frac{1}{(x - n\pi)^2}$$

for all real noninteger x .

If we seek analogues in the complex plane, we are led to functions of the type

$$F(z) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \frac{1}{(z - n - mi)^3}.$$

Observe that, while the real function $\operatorname{cosec}^2 x$ satisfies $\operatorname{cosec}^2(x + \pi) = \operatorname{cosec}^2(x)$ and is periodic with period π , the complex function F just defined satisfies

$$F(z + 1) = F(z), \quad F(z + i) = F(z)$$

and is *doubly periodic* with periods 1 and i . Functions like F are called *elliptic functions* and have a theory that parallels that of the TRIGONOMETRIC FUNCTIONS [III.94].

The function $E(x) = (2\pi)^{-1/2} e^{-x^2/2}$ is called the *Gaussian* (see [III.73 §5]) or *normal* function and appears in probability and the study of diffusion processes. The partial differential equation

$$\frac{\partial^2 \phi}{\partial x^2}(x, t) = K \frac{\partial \phi}{\partial t}(x, t)$$

with x corresponding to distance and t to time provides a reasonable model for diffusion. It is easy to check that $\phi(x, t) = \psi(x, t) = (Kt)^{-1/2} E(x(Kt)^{-1/2})$ is a solution. By sketching a graph of $\psi(x, t)$ as a function of x for various values of t , readers will see that ψ can be considered as the response to a disturbance at $x = 0$ when $t = 0$. By considering the behavior of $\psi(x, t)$ as a function of t for a given value of x , they will see that “the effect at x of a disturbance at the origin becomes noticeable only after a time of the order $x^{1/2}$.” Living cells depend on diffusion processes and the result just given suggests (correctly) that such processes are very slow over long distances. It is plausible that this sets a limit on the size of a single cell: a large organism must be multi-celled.

Statisticians constantly use the related *error function*

$$\operatorname{erf}(x) = \frac{2}{\pi^{1/2}} \int_0^x \exp(-t^2) dt.$$

There is a famous theorem of LIOUVILLE [VI.39] that shows that $\operatorname{erf}(x)$ cannot be expressed as a composition of elementary functions (such as quotients of polynomials, trigonometric functions, and EXPONENTIAL FUNCTIONS [III.25]).

We have been able to look at only a few properties of a few special functions in this article, but even this small sample shows how much interesting mathematics arises when we study one function or a class of particular functions rather than functions in general.

III.88 The Spectrum

G. R. Allan

In the theory of LINEAR MAPS [I.3 §4.2], or *operators*, on a VECTOR SPACE [I.3 §2.3], the notions of EIGENVALUE AND EIGENVECTOR [I.3 §4.3] play an important role. Recall that if V is a vector space (over \mathbb{R} or \mathbb{C}) and if $T : V \rightarrow V$ is a linear mapping, then an *eigenvector* of T is a nonzero vector e in V such that $T(e) = \lambda e$ for some scalar λ ; then λ is the *eigenvalue* corresponding to the eigenvector e . If V is finite dimensional, then the eigenvalues are also the roots of the *characteristic polynomial* $\chi(t) = \det(tI - T)$ of T . Because every nonconstant complex polynomial has a root (the so-called FUNDAMENTAL THEOREM OF ALGEBRA [V.15]), it follows that every linear operator on a finite-dimensional, complex vector space has at least one eigenvalue. If the scalar field is \mathbb{R} , then not all operators have eigenvectors (e.g., consider a rotation about the origin in \mathbb{R}^2).

The linear operators that arise in analysis usually act on infinite-dimensional spaces (see [III.52]). We consider *continuous linear operators* acting on a complex BANACH SPACE [III.64]; these will be referred to simply as *operators* (even though not all linear operators on an infinite-dimensional Banach space are continuous). We shall now see that, for X infinite dimensional, not every such operator has an eigenvalue.

Example 1. Let X be the Banach space $C[0, 1]$, consisting of all continuous, complex-valued functions on the closed interval $[0, 1]$ of the real line. The vector-space structure is the “natural” one (e.g., for $f, g \in X$ the sum $f + g$ is defined by setting $(f + g)(t) = f(t) + g(t)$ for each t and the norm is the *supremum norm*, that is, the largest value of any $|f(t)|$).

Now let u be a continuous complex-valued function on $[0, 1]$. We can associate with it a *multiplication operator* M_u on $C[0, 1]$ as follows. Given a function f , let $M_u(f)$ be the function that takes t to $u(t)f(t)$. It is clear that M_u is linear and continuous. We shall see that whether M_u has an eigenvalue depends on the choice of u . We consider two simple cases.

- (i) Let u be the constant function $u(t) \equiv k$. Then evidently M_u has the single eigenvalue k and every (nonzero) function f in X is an eigenvector.
- (ii) Let $u(t) = t$ for all t . Suppose that the complex number λ is an eigenvalue of M_u . Then there is some $f \in C[0, 1]$, not identically zero, such that $u(t)f(t) = \lambda f(t)$ and so $(t - \lambda)f(t) = 0$ for all

t . But then $f(t) = 0$ for all $t \neq \lambda$, so that, since f is continuous, $f(t) \equiv 0$, contrary to hypothesis. So, for this choice of u , the operator M_u has no eigenvalue.

Let X be a complex Banach space and let T be an operator on X . Then T is said to be *invertible* if and only if there is some operator S on X for which $ST = TS = I$ (here, ST is the composition of S and T , and I is the identity operator on X). It can be shown that T is invertible if and only if T is both *injective* (i.e., $T(x) = 0$ only for $x = 0$) and *surjective* (i.e., $T(X) = X$). The part here that is not just simple algebra is to show that if T is both injective and surjective, then the linear inverse T^{-1} is a *continuous* operator. A complex number λ is an eigenvalue of T precisely if $T - \lambda I$ is *not* injective.

If V is a *finite-dimensional* space, then an injective operator $T : V \rightarrow V$ is necessarily also surjective, and hence invertible. For X infinite dimensional this implication is no longer valid.

Example 2. Let H be the HILBERT SPACE [III.37] ℓ^2 that consists of all sequences $(\xi_n)_{n \geq 1}$ of complex numbers such that $\sum_{n \geq 1} |\xi_n|^2 < \infty$. Let S be the “right-shift” operator defined by $S(\xi_1, \xi_2, \xi_3, \dots) = (0, \xi_1, \xi_2, \dots)$. Then S is injective but not surjective. The “reverse shift” S^* , defined by $S^*(\xi_1, \xi_2, \xi_3, \dots) = (\xi_2, \xi_3, \dots)$, is surjective but not injective.

With this example in mind, we make the following definition.

Definition 3. Let X be a complex Banach space and let T be an operator on X . The *spectrum* of T , denoted by $\text{Sp } T$ (or $\sigma(T)$), is the set of all complex numbers λ such that $T - \lambda I$ is not invertible.

The following remarks should be clear.

- (i) If X is finite dimensional, then $\text{Sp } T$ is just the set of eigenvalues of T .
- (ii) For general X , $\text{Sp } T$ includes the set of eigenvalues of T , but may be larger (e.g., in example 2, 0 is not an eigenvalue of S , but 0 does belong to the spectrum of S).

It is easy to show that the spectrum is always a *bounded* and *closed* (i.e., COMPACT [III.9]) subset of \mathbb{C} . A rather deeper fact is that it is never empty: that is, there will always be some λ for which $T - \lambda I$ is not invertible. That is proved by applying LIOUVILLE’S THEOREM [I.3 §5.6] to the analytic operator-valued function $\lambda \mapsto (\lambda I - T)^{-1}$, defined for λ not in the spectrum of T .

Example 1 continued. We have already seen that not all multiplication operators have eigenvalues. However, they do have an easily described spectrum. Let M_u be such an operator and let S be the set of all values $u(t)$ taken by the function u . Let $\mu = u(t_0)$ be one of these values and consider the operator $M_u - \mu I$. Given any function f in $C[0, 1]$, the value of $(M_u - \mu I)f$ at t_0 is $u(t_0)f(t_0) - \mu f(t_0) = 0$. It follows that $M_u - \mu I$ is not surjective (for instance, the range of $M_u - \mu I$ does not contain any nonzero constant function) and therefore μ belongs to the spectrum of M_u . Thus S is contained in the spectrum of M_u ; it is not hard to show that the two are in fact equal.

We may easily generalize this example to show that if K is *any* nonempty compact subset of \mathbb{C} , then there is a linear operator T with K as its spectrum. Let X be the space of continuous complex-valued functions defined on K , for each $z \in K$, let $u(z) = z$, and let T be the multiplication operator M_u , defined as it was when K was the set $[0, 1]$.

The spectrum is central to most aspects of operator theory. We shall briefly mention a result about Hilbert-space operators, known as the spectral theorem (there are a number of variations).

Let H be a Hilbert space with inner product $\langle x, y \rangle$. A continuous linear operator T on H is called *Hermitian* if $\langle Tx, y \rangle = \langle x, Ty \rangle$ ($x, y \in H$).

Examples 4.

- (i) If H is finite dimensional, then a linear operator S on H is Hermitian if and only if, with respect to some (and hence *every*) ORTHONORMAL BASIS [III.37], S is represented by a Hermitian matrix (i.e., a matrix A with $A = \bar{A}^T$).
- (ii) On the Hilbert space $L_2[0, 1]$, let M_u be the operator of multiplication by a continuous function u (just as in example 1, but now we apply M_u to functions in $L_2[0, 1]$ rather than just $C[0, 1]$). Then M_u is Hermitian if and only if u is real-valued.

If H is finite dimensional and T is a Hermitian operator on H , then H has an orthonormal basis consisting of eigenvectors of T (a “diagonal basis”). Equivalently, $T = \sum_{j=1}^k \lambda_j P_j$, where $\{\lambda_1, \dots, \lambda_k\}$ are the *distinct* eigenvalues of T and P_j is the orthogonal projection of H onto the eigenspace $E_j \equiv \{x \in H : Tx = \lambda_j x\}$.

If H is infinite dimensional and T is a Hermitian operator on H , then it is *not* generally true that H has a

PUP: again, Tim and I think this is OK as ‘Examples’. OK?

PUP: Tim and I would prefer to keep the current numbering system in this article (it will be familiar to readers of maths journals), but let me know if you would strongly like a change.

basis of eigenvectors. But, very importantly, the representation $T = \sum \lambda_j P_j$ does generalize to a representation $T = \int \lambda dP$, a kind of integral with respect to a “projection-valued measure” on the spectrum of T .

There is an intermediate case, for so-called *compact Hermitian operators*, “compactness” being a kind of strong continuity, of great importance in applications. The technicalities are much simpler than in the general case, involving an infinite sum, rather than an integral. A very readable introduction may be found in Young (1988).

Further Reading

Young, N. 1988. *An Introduction to Hilbert Space*. Cambridge: Cambridge University Press.

III.89 Spherical Harmonics

The starting point for FOURIER ANALYSIS [III.27] is the observation that a wide class of periodic functions $f(\theta)$ with period 2π can be decomposed as infinite linear combinations of the TRIGONOMETRIC FUNCTIONS [III.94] $\sin n\theta$ and $\cos n\theta$, or, equivalently, as sums of the form $\sum_{n=-\infty}^{\infty} a_n e^{in\theta}$.

A useful way to think of a periodic function f defined on the real line is as an equivalent function F defined on \mathbb{T} , the unit circle in the complex plane. A typical point on the circle has the form $e^{i\theta}$, and we define $F(e^{i\theta})$ to be $f(\theta)$. (Note that if we add 2π to θ then $F(e^{i\theta})$ does not change because $e^{i\theta} = e^{i(\theta+2\pi)}$ and $f(\theta)$ does not change because f is periodic with period 2π .)

If $f(\theta) = \sum_{n=-\infty}^{\infty} a_n e^{in\theta}$ and $z = e^{i\theta}$, then $F(z) = \sum_{n=-\infty}^{\infty} a_n z^n$. Therefore, if we consider functions defined on \mathbb{T} rather than periodic functions defined on \mathbb{R} , then Fourier analysis decomposes our functions into infinite linear combinations of the functions z^n , where n can be any integer.

What is special about the functions z^n ? The answer is that they are the *characters* of \mathbb{T} , which means that they are the only nonzero continuous complex-valued functions defined on \mathbb{T} that satisfy the relation $\phi(zw) = \phi(z)\phi(w)$ for every z and w in \mathbb{T} .

Now imagine that F is a function defined not on \mathbb{T} but on the two-dimensional set S_2 , which is the unit sphere in \mathbb{R}^3 (defined as the set of points (x, y, z) such that $x^2 + y^2 + z^2 = 1$). More generally, how about functions F defined on S_{d-1} (defined as the set of points (x_1, \dots, x_d) such that $x_1^2 + \dots + x_d^2 = 1$)? Is there a natural way of decomposing such an F , at least if it is suffi-

ciently nice? That is, is there a good way of generalizing Fourier analysis to higher-dimensional spheres?

There is an important and initially discouraging difference between the sphere S_2 and the circle $S_1 = \mathbb{T}$. We defined \mathbb{T} as a set of complex numbers rather than as a set of points in the plane \mathbb{R}^2 because that way it forms a multiplicative group. The sphere, by contrast, does not have a useful group structure (for a clue about why, see QUATERNIONS, OCTONIONS, AND NORMED DIVISION ALGEBRAS [III.78]), so we cannot talk about characters. This makes it less obvious what the “nice” functions should be, into which we might hope to decompose more general functions.

However, there is another way of explaining why the trigonometric functions arise naturally, one that does not involve complex numbers. We can write a typical point in S_1 as (x, y) with $x^2 + y^2 = 1$, or equivalently as $(\cos \theta, \sin \theta)$ for some real number θ . Then our basic functions, if we wish to avoid complex numbers, are $\cos n\theta$ and $\sin n\theta$, but these can also be written in terms of x and y . For instance, $\cos \theta$ and $\sin \theta$ are x and y , respectively, $\cos 2\theta = \cos^2 \theta - \sin^2 \theta = x^2 - y^2$, and so on. (Note that $x^2 - y^2 = 2x^2 - 1 = 1 - 2y^2$, since $x^2 + y^2 = 1$.) In general, $\cos n\theta$ and $\sin n\theta$ can always be written as polynomials in $\cos \theta$ and $\sin \theta$, so the basic trigonometric functions can be thought of as restrictions to the unit circle of certain polynomials.

What are these polynomials? It turns out that they are *harmonic* and *homogeneous*. A harmonic polynomial $p(x, y)$ is one that satisfies the LAPLACE EQUATION [I.3 §5.4] $\Delta p = 0$, where Δp stands for

$$\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2}.$$

For instance, if $p(x, y) = x^2 - y^2$, then $\partial^2 p / \partial x^2 = 2$ and $\partial^2 p / \partial y^2 = -2$, so $x^2 - y^2$ is, as we would hope, a harmonic polynomial. Since the Laplacian Δ is a linear operator, the harmonic polynomials form a vector space. A homogeneous polynomial of degree n is one in which the total degree of each term is n , or equivalently a polynomial $p(x, y)$ such that $p(\lambda x, \lambda y)$ is always equal to $\lambda^n p(x, y)$. For example, $x^3 - 3xy^2$ is homogeneous of degree 3 (and also harmonic). The homogeneous harmonic polynomials of degree n form a subspace of the space of all harmonic polynomials. It has dimension 1 when $n = 0$ and 2 when $n > 0$. (When $n > 0$ it corresponds to the space of functions of the form $A \cos n\theta + B \sin n\theta$. The polynomial $x^3 - 3xy^2$, for instance, corresponds to the function $\cos 3\theta$.)

The notion of a harmonic polynomial generalizes very easily to higher dimensions. For example, in three

dimensions a harmonic polynomial is a polynomial $p(x, y, z)$ such that

$$\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} + \frac{\partial^2 p}{\partial z^2} = 0.$$

A *spherical harmonic* of order n and dimension d is the restriction to the sphere S_{d-1} of a harmonic polynomial in d variables that is homogeneous of degree n .

Here are some of the properties of spherical harmonics that make them particularly useful and closely analogous to the trigonometric polynomials on the circle. We shall fix a dimension d and use the notation $d\mu$ to denote *Haar measure* on the unit sphere $S = S_{d-1}$. Basically, this means that if f is an integrable function from S to \mathbb{R} , then $\int_S f(x) d\mu$ is its average.

(i) Orthogonality. If p and q are spherical harmonics of dimension d and different degrees, then $\int_S p(x)q(x) d\mu = 0$.

(ii) Completeness. Every function $f : S \rightarrow \mathbb{R}$ that belongs to $L^2(S, \mu)$ (meaning that $\int_S |f(x)|^2 d\mu$ exists and is finite) can be written as a sum $\sum_{n=0}^{\infty} H_n$ (with convergence in $L^2(S, \mu)$), where H_n is a spherical harmonic of order n .

(iii) Finite-dimensionality of decomposition. For each d and n , the vector space of spherical harmonics of dimension d and order n is finite dimensional.

From these three properties it follows easily that $L^2(S, \mu)$ has an ORTHONORMAL BASIS [III.37] consisting of spherical harmonics.

Why are spherical harmonics natural, and why are they useful? Both questions can be given several answers: here is one for each.

The Laplace operator Δ , which operates on functions defined on \mathbb{R}^n , can be generalized to functions defined on any RIEMANNIAN MANIFOLD [I.3 §6.10] M . The generalization, denoted Δ_M , is called the *Laplace-Beltrami operator* for M , and its behavior gives one a great deal of information about the geometry of M . In particular, the Laplace-Beltrami operator can be defined for the sphere S_{d-1} , where it is called simply the *Beltrami operator*. It turns out that the spherical harmonics are the EIGENVECTORS [I.3 §4.3] of the Beltrami operator. More precisely, a spherical harmonic of dimension d and order n is an eigenvector with eigenvalue $-n(n + d - 2)$. (Notice that the second derivative of $\cos n\theta$ is $-n^2 \cos n\theta$, which corresponds to the case $d = 2$.) This gives an alternative, more natural (but less elementary) definition of spherical harmonics. This definition, combined with the fact that the Laplace opera-

tor is self-adjoint, explains many of the important properties of spherical harmonics. (See LINEAR OPERATORS AND THEIR PROPERTIES [III.52 §3] for an amplification of this remark.)

One reason for the importance of Fourier analysis is that many important linear operators become diagonal, and hence particularly easy to understand, when they are applied to the Fourier transform of a function. For example, if f is a smooth periodic function and we write it as $\sum_{n \in \mathbb{Z}} a_n e^{in\theta}$, then its derivative is $\sum_{n \in \mathbb{Z}} na_n e^{in\theta}$. Writing $\hat{f}(n)$ for the n th Fourier coefficient of f , we deduce that $\widehat{f'}(n) = n\hat{f}(n)$, which tells us that to differentiate a function f all we have to do is multiply its Fourier transform pointwise by the function $g(n) = n$. This provides a very useful technique for solving differential equations.

As has already been mentioned, spherical harmonics are eigenvalues of the Laplacian, but they also diagonalize several other linear operators. A good example is the *spherical Radon transform*, which is defined as follows. If f is a function from S_{d-1} to \mathbb{R} , then its spherical Radon transform Rf is another function from S_{d-1} to \mathbb{R} , and the value of Rf at a point x is the average value of f over all points y that are orthogonal to x . This is closely related to the more usual Radon transform, which replaces a function defined on the plane by its averages over lines; inverting the Radon transform is important for creating images from the outputs of medical scanners. The spherical harmonics turn out to be eigenfunctions for the spherical Radon transform. More generally, any transform T of the form $Tf(x) = \int_S w(x \cdot y) f(y) d\mu(y)$, where w is a suitable function (or generalized function), is diagonalized by spherical harmonics. The eigenvalue associated with a given spherical harmonic can be calculated by the so-called *Funk-Hecke formula*.

Spherical harmonics give a way of linking CHEBYSHEV AND LEGENDRE POLYNOMIALS [III.87], and showing that both of them are natural concepts. The Chebyshev polynomials are those polynomials in x that are also spherical harmonics of dimension 2: that is, that are equal on S_1 to homogeneous harmonic polynomials in two variables. For instance, because $x^2 + y^2 = 1$ for every (x, y) in the circle S_1 , the function $x^3 - 3xy^2$ that we considered earlier is equal on S_1 to the function $4x^3 - 3x$, so $4x^3 - 3x$ is a Chebyshev polynomial. The Legendre polynomials are those polynomials in x that are equal to spherical harmonics of dimension 3. For example, if $p(x, y, z) = 2x^2 - y^2 - z^2$ then $\Delta p = 0$, and $p(x, y, z) = 3x^2 - 1$ everywhere on S_2 ,

since $x^2 + y^2 + z^2 = 1$. Therefore, $3x^2 - 1$ is a Legendre polynomial.

Here is a sketch of a proof that these polynomials are equal to the Chebyshev and Legendre polynomials as they are usually defined. The usual definition is that they are sequences of polynomials, one for each degree, that are uniquely determined by certain orthogonality relations. Because spherical harmonics of different orders are orthogonal, the polynomials just described also satisfy certain orthogonality relations. When one works out what these are, one discovers that they are precisely the relations that define the Chebyshev and Legendre polynomials.

III.90 Symplectic Manifolds

Gabriel P. Paternain

Symplectic geometry is the geometry that governs classical physics, and more generally plays an important role in helping us to understand the actions of groups on manifolds. It shares some features with Riemannian geometry and complex geometry, and there is an important special class of manifolds, the *Kähler manifolds*, in which all three geometric structures are unified.

1 Symplectic Linear Algebra

Just as RIEMANNIAN GEOMETRY [I.3 §6.10] is based on EUCLIDEAN GEOMETRY [I.3 §6.2], symplectic geometry is based on the geometry of the so-called *linear symplectic space* $(\mathbb{R}^{2n}, \omega_0)$.

Given two vectors $v = (q, p)$ and $v' = (q', p')$ in the plane \mathbb{R}^2 , the *signed area* $\omega_0(v, v')$ of the parallelogram spanned by v and v' is given by the formula

$$\omega_0(v, v') = \det \begin{pmatrix} q' & q \\ p' & p \end{pmatrix} = pq' - qp'.$$

It can also be written using matrices and inner products as $\omega_0(v, v') = v' \cdot Jv$, where J is the 2×2 matrix

$$J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

If a linear transformation $A : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is area preserving and orientation preserving, then $\omega_0(Av, Av') = \omega_0(v, v')$ for every v and v' .

Symplectic geometry studies two-dimensional signed area measurements like this, as well as transformations that preserve these measurements, but it applies to general spaces of dimension $2n$ rather than just to the plane.

If we split \mathbb{R}^{2n} up as $\mathbb{R}^n \times \mathbb{R}^n$, then we can write a vector v in \mathbb{R}^{2n} as $v = (q, p)$, where q and p each belong to \mathbb{R}^n . The *standard symplectic form* $\omega_0 : \mathbb{R}^{2n} \times \mathbb{R}^{2n} \rightarrow \mathbb{R}$ is defined by the formula

$$\omega_0(v, v') = p \cdot q' - q \cdot p',$$

where “ \cdot ” denotes the usual inner product in \mathbb{R}^n . Geometrically, $\omega_0(v, v')$ can be interpreted as the sum of the signed areas of the parallelograms spanned by the projections of v and v' to the $q_i p_i$ -planes. In terms of matrices, we can write

$$\omega_0(v, v') = v' \cdot Jv, \quad (1)$$

where J is the $2n \times 2n$ matrix

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \quad (2)$$

and I is the $n \times n$ identity matrix.

A linear map $A : \mathbb{R}^{2n} \rightarrow \mathbb{R}^{2n}$ that preserves the product ω_0 of any two vectors (that is, $\omega_0(Av, Av') = \omega_0(v, v')$ for all $v, v' \in \mathbb{R}^{2n}$) is called a *symplectic linear transformation*; equivalently, a $2n \times 2n$ matrix A is symplectic if and only if $A^T J A = J$, where A^T is the transpose of A . Symplectic linear transformations are to symplectic geometry as rigid motions are to Euclidean geometry. The set of all symplectic linear transformations of $(\mathbb{R}^{2n}, \omega_0)$ is one of the classical LIE GROUPS [III.50 §1] and is denoted by $\text{Sp}(2n)$. One can show that symplectic matrices $A \in \text{Sp}(2n)$ always have DETERMINANT [III.15] 1, and are thus volume preserving. However, the converse does not hold when $n \geq 2$. For instance, if $n = 2$, the linear map

$$(q_1, q_2, p_1, p_2) \mapsto (aq_1, q_2/a, ap_1, p_2/a)$$

has determinant 1 for any $a \neq 0$, but it is symplectic only if $a^2 = 1$.

The standard symplectic form ω_0 has three properties worth noting. First, it is *bilinear*: the expression $\omega_0(v, v')$ varies linearly in v when v' is held fixed, and vice versa. Second, it is *antisymmetric*: we have $\omega_0(v, v') = -\omega_0(v', v)$ for all v and v' , and in particular $\omega_0(v, v) = 0$. Finally, it is *nondegenerate*, which means that for every nonzero v there is a nonzero v' such that $\omega_0(v, v') \neq 0$. The standard symplectic form ω_0 is not the only form that obeys these three properties; however, it turns out that any form with these three properties can be converted into the standard form ω_0 after an invertible linear change of variables. (This is a special case of *Darboux's theorem*.) Thus $(\mathbb{R}^{2n}, \omega_0)$ is essentially the “only” linear symplectic geometry in $2n$ dimensions. There are no symplectic forms in odd-dimensional spaces.

2 Symplectic Diffeomorphisms of $(\mathbb{R}^{2n}, \omega_0)$

In Euclidean geometry, all rigid motions are automatically linear (or affine) transformations. However, in symplectic geometry there are many more symplectic maps than just the symplectic linear transformations. These nonlinear symplectic maps in $(\mathbb{R}^{2n}, \omega_0)$ are one of the principal objects of study in symplectic geometry.

Let $U \subset \mathbb{R}^{2n}$ be an open set. Recall that a map $\phi : U \rightarrow \mathbb{R}^{2n}$ is called *smooth* if it has continuous partial derivatives of all orders. A *diffeomorphism* is a smooth map with smooth inverse.

A smooth nonlinear map $\phi : U \rightarrow \mathbb{R}^{2n}$ is said to be *symplectic* if, for every $x \in U$, the *Jacobian matrix* $\phi'(x)$ of first derivatives of ϕ is a symplectic linear transformation. Informally, a symplectic map is one that behaves like a symplectic linear transformation at infinitesimally small scales. Since symplectic linear transformations have determinant 1, we can conclude using several-variable calculus that a symplectic map is always locally volume preserving and locally invertible; roughly speaking, this means that the map $\phi : A \rightarrow \phi(A)$ is invertible whenever A is a sufficiently small subset of U , and $\phi(A)$ has the same volume as A . However, the converse is not true when $n \geq 2$; the class of symplectic maps is much more restricted than that of volume-preserving maps. In fact, Gromov's non-squeezing theorem (see below) shows how striking this difference can be.

Symplectic maps have been around for quite a long time in Hamiltonian mechanics under the name of *canonical transformations*. We briefly explain this in the next subsection.

2.1 Hamilton's Equations

How can we produce nonlinear symplectic maps? Let us begin by exploring a familiar example. Consider the motion of a simple pendulum with length l and mass m and let $q(t)$ be the angle it makes with the vertical at time t . The equation of motion is

$$\frac{d^2 q}{dt^2} + \frac{g}{l} \sin q = 0,$$

where g is the acceleration due to gravity. If we define the *momentum* p as $p = ml^2 \dot{q}$, then we may transform this second-order differential equation into a first-order system in the *phase plane* \mathbb{R}^2 , namely

$$\frac{d}{dt}(q, p) = X(q, p), \quad (3)$$

where the *vector field* $X : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is given by the formula $X(q, p) = (p/ml^2, -mgl \sin q)$. For each $(q(0), p(0)) \in \mathbb{R}^2$ there is a unique solution $(q(t), p(t))$ to (3) with initial condition $(q(0), p(0))$. Then for any fixed time t we obtain an *evolution map* (or *flow*) $\phi_t : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ given by $\phi_t(q(0), p(0)) = (q(t), p(t))$, which has the remarkable property of being *area preserving*. This can be deduced from the observation that X is *divergence free*, or in other words that

$$\frac{d}{dq} \frac{p}{ml^2} + \frac{d}{dp} (-mgl \sin q) = 0.$$

In fact, for every time t , ϕ_t is a symplectic map on (\mathbb{R}^2, ω_0) .

More generally, any system in classical mechanics with finitely many degrees of freedom can be reformulated in a similar fashion, so that the evolution maps ϕ_t are always symplectic maps; in this context, they are also known as *canonical transformations*. The Irish mathematician WILLIAM ROWAN HAMILTON [VI.37] showed us how to do this in general more than 170 years ago. Given any smooth function $H : \mathbb{R}^{2n} \rightarrow \mathbb{R}$ (called the *Hamiltonian*), the system of first-order differential equations given by

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}, \quad i = 1, \dots, n, \quad (4)$$

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}, \quad i = 1, \dots, n, \quad (5)$$

will (under some mild growth assumptions on H , which we ignore here) give rise to evolution operators $\phi_t : \mathbb{R}^{2n} \rightarrow \mathbb{R}^{2n}$, which are symplectic maps on $(\mathbb{R}^{2n}, \omega_0)$ for every time t . To see the connection with the symplectic form ω_0 , observe that we may rewrite (4) and (5) in the following equivalent form:

$$\frac{dx}{dt} = J \nabla H(x), \quad (6)$$

where ∇H is the usual GRADIENT [I.3 §5.3] of H and J was defined in (2). From (6), (1), and the antisymmetry property of ω_0 , it is then not difficult to verify that ϕ_t is a diffeomorphism for every t (the main trick is to compute the derivative of $\omega_0(\phi'_t(x)v, \phi'_t(x)v')$ in t and check that it equals zero).

We have already pointed out that symplectic maps are volume preserving. The preservation of volume by Hamiltonian systems (a result known as *Liouville's theorem*) attracted considerable attention in the nineteenth century and it was a driving force in the development of ERGODIC THEORY [V.11], which studies recurrence properties of measure-preserving transformations.

Symplectic maps or canonical transformations play an important role in classical physics, as they allow one to replace a complicated system by an equivalent system that is simpler to analyze.

2.2 Gromov's Nonsqueezing Theorem

What is the difference between a symplectic map and a volume-preserving map? In order to answer this question, suppose that we have two connected open sets U and V in \mathbb{R}^{2n} and that we wish to embed one into the other using a symplectic map. This means that we are looking for a symplectic map $\phi : U \rightarrow V$ such that ϕ is a homeomorphism onto its image. We know such a ϕ must be volume preserving, so we clearly have the restriction that the volume of U should be at most the volume of V , but is this restriction all that matters? Consider the open ball $B(R) = \{x \in \mathbb{R}^{2n} : |x| < R\}$, which has radius R and center at the origin, and clearly has finite volume. It is not hard to embed it symplectically into the infinite-volume cylinder given by

$$C(r) = \{(q, p) \in \mathbb{R}^{2n} : q_1^2 + q_2^2 < r^2\}$$

for any positive R and r . Indeed, the linear symplectic map

$$(q, p) \mapsto (aq_1, aq_2, q_3, \dots, q_n, p_1/a, p_2/a, p_3, \dots, p_n)$$

will do the trick when a is sufficiently small and positive. However, the situation is radically different if instead we consider the infinite-volume cylinder

$$Z(r) = \{(q, p) \in \mathbb{R}^{2n} : q_1^2 + p_1^2 < r^2\}.$$

We could try with a similar linear map like

$$(q, p) \mapsto (aq_1, q_2/a, q_3, \dots, q_n, ap_1, p_2/a, p_3, \dots, p_n).$$

This map is volume preserving (it has determinant 1) and for a small it embeds $B(R)$ into $Z(r)$. However, it is symplectic only if $a = 1$, so it will give a symplectic embedding only if $R \leq r$. One is tempted to think that if $R > r$, then there should still be a nonlinear symplectic embedding squeezing $B(R)$ into $Z(r)$, but a remarkable theorem of Gromov from 1985 asserts that it is not possible to find such a map.

In spite of this deep result of Gromov, and other results that followed it, we still do not know much about how sets in \mathbb{R}^{2n} embed into one another.

3 Symplectic Manifolds

Recall from DIFFERENTIAL TOPOLOGY [IV.9] that a *manifold* of dimension d is a TOPOLOGICAL SPACE [III.92]

such that each point has a neighborhood that is homeomorphic to an open set in Euclidean space \mathbb{R}^d . One can think of \mathbb{R}^d as a *local model* for this manifold, in the sense that it describes what the manifold looks like at very small distance scales. Recall also that a *smooth* manifold is one for which the “transition functions” are smooth. This means that if $\psi : U \rightarrow \mathbb{R}^d$ and $\varphi : V \rightarrow \mathbb{R}^d$ are coordinate charts, then the transition function $\psi \circ \varphi^{-1}$ between the open sets $\phi(U \cap V)$ and $\psi(U \cap V)$ is smooth.

A symplectic manifold is defined similarly, but now the local model is the linear symplectic space $(\mathbb{R}^{2n}, \omega_0)$. More precisely, a symplectic manifold M is a manifold of dimension $2n$ that can be covered with domains of coordinate charts whose transition functions are symplectic diffeomorphisms of $(\mathbb{R}^{2n}, \omega_0)$.

Of course, any open set in $(\mathbb{R}^{2n}, \omega_0)$ is a symplectic manifold. An example of a compact symplectic manifold is the torus \mathbb{T}^{2n} , which is obtained as the quotient of \mathbb{R}^{2n} by the action of \mathbb{Z}^{2n} . In other words, two points $x, y \in \mathbb{R}^{2n}$ are equivalent if $x - y$ has integer coordinates. Other important examples of symplectic manifolds include RIEMANN SURFACES [III.81], COMPLEX PROJECTIVE SPACE [III.74], and COTANGENT BUNDLES [IV.10 §5]. However, it is a wide open problem to determine, given a compact manifold, whether it can be assigned a system of coordinate charts that makes it symplectic.

We have seen that in $(\mathbb{R}^{2n}, \omega_0)$, one can assign an “area” $\omega_0(v, v')$ to any parallelogram in the space \mathbb{R}^{2n} . In a symplectic manifold M , one can similarly assign an area $\omega_p(v, v')$, but only to *infinitesimal* parallelograms based at a point $p \in M$. The axes of such a parallelogram are two infinitesimal vectors (or more precisely *tangent vectors*) v and v' . There is a unique way to do this so that all the coordinate charts for M are symplectic diffeomorphisms. In the language of DIFFERENTIAL FORMS [III.16], the map $p \mapsto \omega_p$ is an antisymmetric nondegenerate 2-form, which can then be used to compute the “area” $\int_S \omega$ of noninfinitesimal two-dimensional surfaces S in M . One can show that for any sufficiently small closed surface S , the integral $\int_S \omega$ vanishes, so ω is a *closed* form. Indeed, one can define a symplectic manifold more abstractly (without reference to charts) as a smooth manifold equipped with a closed, antisymmetric nondegenerate 2-form ω ; a classical theorem of Darboux asserts that this abstract definition is equivalent to the more concrete definition using coordinate charts.

PUP: difference between this display and the one after next is indeed deliberately different.

Finally, a special class of symplectic manifolds is given by *Kähler manifolds*. These are symplectic manifolds that are also *complex manifolds*, in such a way that the two structures are naturally compatible, a condition that generalizes the relationship (1). Observe that if one identifies points (q, p) in \mathbb{R}^{2n} with points $p + iq$ in \mathbb{C}^n , then the linear transformation $J : \mathbb{R}^{2n} \rightarrow \mathbb{R}^{2n}$ becomes the operation of multiplication by i :

$$J : (z_1, \dots, z_n) \mapsto (iz_1, \dots, iz_n).$$

Thus the identity (1) relates the symplectic structure (as given by ω_0), the complex structure (as given by J), and the Riemannian structure (as given by the dot product “ \cdot ”). A *complex manifold* is a manifold that at small distance scales looks like regions of \mathbb{C}^n , with the transition functions required to be *holomorphic* [I.3 §5.6]. (A smooth map $f : U \subset \mathbb{C}^n \rightarrow \mathbb{C}^n$ is said to be holomorphic if each coordinate component of $f(z_1, \dots, z_n)$ is holomorphic in each variable z_k .) On a complex manifold we can multiply tangent vectors by i . This gives us at each point $p \in M$ a linear map J_p such that $J_p^2 v = -v$ for all tangent vectors v at p . A Kähler manifold is a complex manifold M with a symplectic structure ω (which computes signed areas of infinitesimal parallelograms) and a Riemannian metric g (which computes an inner product $g_p(v, v')$ of any two tangent vectors v, v' at p); these two structures are linked together by the analogue of (1), namely

$$\omega_p(v, v') = g_p(v', J_p v).$$

Examples of Kähler manifolds include complex vector spaces \mathbb{C}^n , Riemann surfaces, and complex projective spaces \mathbb{CP}^n .

An example of a compact symplectic manifold that is not Kähler can be obtained by taking the quotient of \mathbb{R}^4 by a symplectic action of a group that looks like \mathbb{Z}^4 but with a group operation that differs from the usual one. The change in the group structure manifests itself as a topological property (an odd first Betti number) that prevents the quotient being Kähler.

Further Reading

- Arnold, V. I. 1989. *Mathematical Methods of Classical Mechanics*, 2nd edn. Graduate Texts in Mathematics, volume 60. New York: Springer.
- McDuff, D., and D. Salamon. 1998. *Introduction to Symplectic Topology*, 2nd edn. Oxford Mathematical Monographs. Oxford: Clarendon Press/Oxford University Press.

III.91 Tensor Products

If U, V , and W are VECTOR SPACES [I.3 §2.3] over some field, then a *bilinear map* from $U \times V$ to W is a map ϕ obeying the rules

$$\phi(\lambda u + \mu u', v) = \lambda \phi(u, v) + \mu \phi(u', v)$$

and

$$\phi(u, \lambda v + \mu v') = \lambda \phi(u, v) + \mu \phi(u, v').$$

That is, it is linear in each variable separately.

Many important maps, such as INNER PRODUCTS [III.37], are bilinear. The *tensor product* $U \otimes V$ of two vector spaces U and V is a way of capturing the idea of the “most general” bilinear map that we can define on $U \times V$. To get an idea of what this might mean, let us try to imagine a “completely arbitrary” bilinear map from $U \times V$ to a “completely arbitrary” vector space W , and let us use the notation $u \otimes v$ instead of $\phi(u, v)$. Now because our linear map is perfectly general, all we know about it is what we can deduce from the fact that it is bilinear. For example, we know that $u \otimes v_1 + u \otimes v_2 = u \otimes (v_1 + v_2)$. This example might suggest that all elements of $U \otimes V$ are of the form $u \otimes v$, but that is certainly not the case: for instance, in general there is no way of simplifying an expression such as $u_1 \otimes v_1 + u_2 \otimes v_2$. (This reflects the fact that the set of values taken by a bilinear map from $U \times V$ to W is not in general a subspace of W .)

Thus, a typical element of $U \otimes V$ is a *linear combination* of elements of the form $u \otimes v$, with the rule that different linear combinations give the same element of $U \otimes V$ whenever they are forced to by the bilinearity property: for instance, $(u_1 + 2u_2) \otimes (v_1 - v_2)$ will always be equal to

$$u_1 \otimes v_1 + 2u_2 \otimes v_1 - u_1 \otimes v_2 - 2u_2 \otimes v_2.$$

A more formal way of expressing the above ideas is to say that $U \otimes V$ has a *universal property*. (See GROUPS AND GEOMETRY [IV.11] for some other examples of universal properties. See also CATEGORIES [III.8].) The property in question is the following: given any bilinear map ϕ from $U \times V$ to a space W , we can find a *linear* map α from $U \otimes V$ to W such that $\phi(u, v) = \alpha(u \otimes v)$ for every u and v . That is, every bilinear map ϕ defined on $U \times V$ is naturally associated with a linear map defined on $U \otimes V$. (This linear map takes $u \otimes v$ to $\phi(u, v)$: the identifications made in the definition of the tensor product ensure that we can extend this to linear combinations of such elements in a consistent way.)

It is not hard to show that if U and V are finite dimensional, with bases u_1, \dots, u_m and v_1, \dots, v_n , then the

vectors $u_i \otimes v_j$ form a basis for $U \otimes V$. Other important properties of the tensor product are that it is commutative and associative, in the sense that $U \otimes V$ is naturally isomorphic to $V \otimes U$ and $U \otimes (V \otimes W)$ is naturally isomorphic to $(U \otimes V) \otimes W$.

We have been discussing tensor products of vector spaces, but the definition can easily be generalized to any algebraic structure for which some notion of bilinearity makes sense, such as a MODULE [III.83 §3] or a C^* -ALGEBRA [IV.19 §3]. Sometimes the tensor product of two structures is not what you would immediately expect. For instance, let \mathbb{Z}_n be the set of integers mod n , and consider both \mathbb{Z}_n and \mathbb{Q} as modules over \mathbb{Z} . Then their tensor product is zero. This reflects the fact that every bilinear map from $\mathbb{Z}_n \times \mathbb{Q}$ must be the zero map.

Tensor products occur in many mathematical contexts. For a good example, see QUANTUM GROUPS [III.77].

Transcendental Numbers

See IRRATIONAL AND TRANSCENDENTAL NUMBERS [III.43]

III.92 Topological Spaces

Ben Green

A topological space is the most basic context in which one can understand the notion of a CONTINUOUS FUNCTION [I.3 §5.2].

Let us recall a standard definition of what it means for a function $f : \mathbb{R} \rightarrow \mathbb{R}$ to be continuous. Suppose that $f(x) = y$. Then f is continuous at x provided that $f(x')$ is close to y whenever x' is close to x . Of course, to make this a mathematically rigorous notion we have to be precise about the meaning of “close.” We could say that $f(x')$ is close to y if $|f(x') - f(x)| < \varepsilon$, where $\varepsilon > 0$ is some small positive constant. And we could deem x to be close to x' whenever $|x - x'| < \delta$, where δ is another positive constant.

We say that f is *continuous at x* if an appropriate δ can be found, regardless of how small ε was chosen to be (δ is allowed, of course, to depend on ε). And f is said to be *continuous* if it is continuous at every point x on the real line.

How might we generalize this notion, replacing \mathbb{R} by an arbitrary set X ? Our existing definition makes sense only if we can decide when two points $x, x' \in X$ are close. For a general set, which might not be nicely

embedded in Euclidean space, this is impossible without the addition of further structure. (When such structure is added one has the notion of a METRIC SPACE [III.58]: metric spaces are less general than topological spaces.)

If the notion of closeness is unavailable, how should one define continuity? The answer may be found in the notion of an *open set*. A set $U \subset \mathbb{R}$ is said to be *open* if for any point x in U there is an interval (a, b) that contains x (that is, $a < x < b$) and is contained in U .

It is an amusing exercise to check that if $f : \mathbb{R} \rightarrow \mathbb{R}$ is continuous, and if U is open, then $f^{-1}(U)$ is open. Conversely, if $f^{-1}(U)$ is open for every open set U , then f is continuous. Thus, at least for functions from \mathbb{R} to \mathbb{R} , one may characterize continuity purely in terms of open sets. The notion of closeness is used only when it comes to defining what an open set is.

We now turn to the formal definition. A *topological space* is a set X together with a collection \mathcal{U} of subsets of X (called the “open sets”) satisfying the following axioms.

- The empty set \emptyset and the set X are both open.
- \mathcal{U} is closed under taking arbitrary unions (so if $(U_i)_{i \in I}$ is a collection of open sets, then so is $\bigcup_{i \in I} U_i$).
- \mathcal{U} is closed under taking finite intersections (so if U_1, \dots, U_k are open sets, then so is $U_1 \cap \dots \cap U_k$).

The collection \mathcal{U} is called a *topology* on X . It is easy to verify that the usual open subsets of \mathbb{R} satisfy the above axioms: thus, \mathbb{R} forms a topological space with these sets.

A subset of a topological space is called *closed* if and only if its complement is open. Note that “closed” does not mean “not open”: for example, in the space \mathbb{R} , the half-open interval $[0, 1)$ is neither open nor closed, and the empty set is both open and closed.

Note that we do not demand many properties from our open sets: this makes the notion of topological space a rather general one. Indeed, under many circumstances the concept is a little *too* general: then it can be convenient to assume that a topological space has further properties. For instance, a topological space X is called *Hausdorff* if, for any two distinct points x_1 and x_2 in X , there are disjoint open sets U_1 and U_2 that contain x_1 and x_2 , respectively. Hausdorff topological spaces (of which \mathbb{R} is an obvious example) have many useful properties that general topological spaces do not necessarily have.

We saw earlier that for functions from \mathbb{R} to \mathbb{R} the notion of continuity could be formulated entirely in terms of open sets. This means that we can define continuity for functions between topological spaces: if X and Y are two topological spaces and if $f : X \rightarrow Y$ is a function between them, then we simply define f to be continuous if $f^{-1}(U)$ is open for every open set $U \subset Y$. Remarkably, we have found a useful definition of continuity that does not rely on a notion of distance.

A continuous map that has a continuous inverse is known as a *homeomorphism*. If there is a homeomorphism between two spaces X and Y , then they are regarded as equivalent from the point of view of topology. In topology texts one will often see it said that a topologist is unable to distinguish between a doughnut and a teacup because each can be continuously deformed into the other (imagine that they are both made of modeling clay).

If X is a topological space, then a very useful way of describing the topology on X is by giving a *basis* for it. This is a subcollection $\mathcal{B} \subseteq \mathcal{U}$ with the property that every open set (that is, every element of \mathcal{U}) is a union of open sets in \mathcal{B} . A basis for \mathbb{R} with the usual topology is the collection of open intervals $\{(a, b) : a < b\}$, and a basis for \mathbb{R}^2 is the collection of *open balls*: that is, sets of the form $\{B_\delta(x) = \{y : |x - y| < \delta\}\}$.

Let us give some examples.

The discrete topology. Let X be any set whatsoever, and take \mathcal{U} to be the collection of all subsets of X . It is a simple matter to check that the axioms for a topological space are satisfied.

Euclidean spaces. Let $X = \mathbb{R}^d$, and let \mathcal{U} contain all sets that are open in the Euclidean metric. That is, $U \subseteq X$ is open if, for every $u \in U$, there is $\delta > 0$ such that $B_\delta(u)$ is contained in U . It is only slightly more taxing to check that the axioms are satisfied in this case. More generally, for any metric space the open sets can be defined in a similar way and they form a topological space.

Subspace topology. If X is a topological space and if $S \subseteq X$, then we may make S a topological space. We declare the open sets in S to be all sets of the form $S \cap U$, where $U \in \mathcal{U}$ is an open set in X .

The Zariski topology. This is used in ALGEBRAIC GEOMETRY [IV.7]. It is specified by giving its closed sets (and hence, by complementation, its open sets)—these are the zero loci of systems of polynomial equations. On

\mathbb{C}^2 , for example, these closed sets are precisely the sets of the form

$$\{(z_1, z_2) : f_1(z_1, z_2) = f_2(z_1, z_2) = \cdots = f_k(z_1, z_2) = 0\},$$

where f_1, \dots, f_k are polynomials. To show that this defines a topology is somewhat nontrivial, the difficulty being to show that an arbitrary intersection of closed sets is closed (which is equivalent to the assertion that an arbitrary union of open sets is open). This is a consequence of Hilbert's basis theorem.

The notion of topological space is a very good example of the power of abstraction in mathematics. The definition is simple and covers a wide variety of natural situations, yet it has enough content that one can make interesting definitions and prove theorems purely within the world of topological spaces. It is often fun to take a familiar concept, that applies to \mathbb{R} or \mathbb{R}^2 , say, and try to find an analogue of it in the world of general topological spaces. We give two examples.

Connectedness. The rough idea of connectedness is that a connected set is one that does not break up into pieces in an obvious way. Most people would imagine that they could discern, from a list of pictures of reasonably sensible subsets of \mathbb{R}^2 , which were connected and which were not. But can one give a precise mathematical definition that applies to all sets, including potentially very wild ones, and says whether they are connected or not? For example, is the space

$$S = ((\mathbb{Q} \times \mathbb{R}) \cup (\mathbb{R} \times \mathbb{Q})) \setminus (\mathbb{Q} \times \mathbb{Q})$$

of lines with exactly one rational coordinate (with the subspace topology) connected or not? It turns out that a definition can indeed be given, and moreover that it applies not just to \mathbb{R}^2 but to general topological spaces. We say that a space X is *connected* if there is no decomposition $X = U_1 \cup U_2$ of X into two disjoint, nonempty, open sets. We leave it to the reader to decide whether S is connected or not.

Compactness. This is one of the most important concepts in all of mathematics, but it can appear strange at first sight. It comes from attempting to abstract the notion of a closed and bounded set (in \mathbb{R}^2 , say) to a general topological space. We say that X is *compact* if, given any collection C of open sets U that cover X (i.e., whose union is X), we may find a finite subcollection $\{U_1, \dots, U_k\} \subseteq C$ that still covers X . Specializing this definition to \mathbb{R}^2 with the usual topology, it can indeed

be proved that a set $S \subseteq \mathbb{R}^2$ is compact (in the subspace topology) if and only if it is closed and bounded. See COMPACTNESS AND COMPACTIFICATION [III.9] for more information.

III.93 Transforms

T. W. Körner

If we have a finite sequence a_0, a_1, \dots, a_n of real numbers (written briefly as \mathbf{a}), then we can look at the polynomial

$$P_{\mathbf{a}}(t) = a_0 + a_1 t + \dots + a_n t^n.$$

Conversely, given a polynomial Q of degree $m \leq n$, we can recover a unique sequence b_0, b_1, \dots, b_n such that

$$P_{\mathbf{b}}(t) = b_0 + b_1 t + \dots + b_n t^n$$

by, for example, taking $b_k = Q^{(k)}(0)/k!$.

We observe that if a_0, a_1, \dots, a_n and b_0, b_1, \dots, b_n are finite sequences with $a_r = b_r = 0$ for $r > \frac{1}{2}n$, then

$$P_{\mathbf{a}}(t)P_{\mathbf{b}}(t) = P_{\mathbf{a} * \mathbf{b}}(t),$$

where $\mathbf{a} * \mathbf{b} = \mathbf{c}$ is a sequence c_0, c_1, \dots, c_{2n} given by

$$c_k = a_0 b_k + a_1 b_{k-1} + \dots + a_k b_0,$$

where we interpret a_i and b_i as 0 if $i > n$. This sequence is called the *convolution* of the sequences \mathbf{a} and \mathbf{b} .

To see the kind of use that one can make of this observation, consider what happens when we throw two dice, the first of which has probability a_u of showing u and the second of which has probability b_v of showing v . The probability that their sum is k is given by the number c_k defined above. If we take both a_u and $b(u)$ to be the probability of throwing u with an ordinary fair die (so they are equal to $\frac{1}{6}$ if $1 \leq u \leq 6$, and 0 otherwise), then

$$\begin{aligned} P_{\mathbf{c}}(t) &= P_{\mathbf{a}}(t)P_{\mathbf{b}}(t) \\ &= \left(\frac{1}{6}(t + t^2 + \dots + t^6)\right)^2. \end{aligned}$$

This polynomial can be rewritten as

$$\begin{aligned} &\frac{1}{36}(t(t+1)(t^4+t^2+1))(t(t^2+t+1)(t^3+1)) \\ &= \frac{1}{36}(t(t+1)(t^2+t+1))(t(t^4+t^2+1)(t^3+1)) \\ &= P_{\mathbf{A}}(t)P_{\mathbf{B}}(t), \end{aligned}$$

where \mathbf{A} and \mathbf{B} are two different sequences, given by $A_1 = A_4 = \frac{1}{6}$, $A_2 = A_3 = \frac{2}{6}$, and $A_u = 0$ otherwise, and $B_1 = B_3 = B_4 = B_5 = B_6 = B_8 = \frac{1}{6}$, and $B_v = 0$ otherwise. Thus, if we take two fair dice A and B and number A so that it has 2 on two faces, 3 on two faces, 1 on one face, and 4 on the remaining face, and we

number B so that it has 1, 3, 4, 5, 6, and 8 on its faces, then the probability of throwing a sum k is the same as with an ordinary pair of dice. It is not hard to show, by considering the roots of the polynomial $t + t^2 + \dots + t^6$, that this is the only nonstandard labeling of dice with strictly positive integers that has this property.

These general ideas are easily extended to infinite sequences. If \mathbf{a} is the sequence a_0, a_1, \dots , we can define an “infinite polynomial” $(\mathcal{G}\mathbf{a})(t)$ to be $\sum_{r=0}^{\infty} a_r t^r$. For the moment, we shall proceed formally, without worrying in what sense the sum exists. Observe that, much as before,

$$(\mathcal{G}\mathbf{a})(t)(\mathcal{G}\mathbf{b})(t) = (\mathcal{G}(\mathbf{a} * \mathbf{b}))(t),$$

where the infinite sequence $\mathbf{c} = \mathbf{a} * \mathbf{b}$ is given by

$$c_k = a_0 b_k + a_1 b_{k-1} + \dots + a_k b_0.$$

(Again, we call this the convolution of \mathbf{a} and \mathbf{b} .)

There is a well-known problem in which we are asked how many ways there are of making change for r units of currency using notes of given denominations. (For example, we can ask how many ways there are of making \$43 out of \$1 and \$5 bills.) If we can make r units in a_r ways using one set of denominations and b_r ways using a completely different set, then it is not hard to see that, if we are allowed to use both sets of denominations, we can make up k units in c_k ways, where c_k is again the number defined earlier.

Let us see how this applies in the simple case where a_r is the number of ways of making up r dollars using \$1 bills and b_r is the number of ways of making up r dollars using \$2 bills. We observe that

$$\begin{aligned} (\mathcal{G}\mathbf{a})(t) &= \sum_{r=0}^{\infty} t^r = \frac{1}{1-t}, \\ (\mathcal{G}\mathbf{b})(t) &= \sum_{r=0}^{\infty} t^{2r} = \frac{1}{1-t^2}, \end{aligned}$$

and so, using partial fractions,

$$\begin{aligned} (\mathcal{G}\mathbf{c})(t) &= (\mathcal{G}(\mathbf{a} * \mathbf{b}))(t) = (\mathcal{G}\mathbf{a})(t)(\mathcal{G}\mathbf{b})(t) \\ &= \frac{1}{(1-t)(1-t^2)} = \frac{1}{(1-t)^2(1+t)} \\ &= \frac{1}{2(1-t)^2} + \frac{1}{4(1+t)} - \frac{1}{4(1-t)} \\ &= \frac{1}{2} \sum_{r=0}^{\infty} (r+1)t^r + \frac{1}{4} \sum_{r=0}^{\infty} (-1)^r t^r - \frac{1}{4} \sum_{r=0}^{\infty} t^r \\ &= \sum_{r=0}^{\infty} \frac{2r+1+(-1)^r}{2} t^r. \end{aligned}$$

Thus we can make change for r dollars in $\frac{1}{2}(r+1)$ ways when r is odd and $\frac{1}{2}(r+2)$ ways when r is even. In this

T&T note: this article could certainly be reduced if and when we get desperate.

simple case it is easy to obtain the result directly but the method indicated works automatically in all cases. (The calculations can be made easier if we allow ourselves to work with complex roots.)

We have produced a “generating function transform” or “ \mathcal{G} -transform,” which takes a sequence a_0, a_1, \dots into a Taylor series $\sum_{r=0}^{\infty} a_r x^r$. (These names are not standard: most mathematicians would simply talk about GENERATING FUNCTIONS [IV.22 §§2.4, 3].) The next two examples show how we can use \mathcal{G} -transforms to restate problems about sequences as problems about Taylor series. Consider first the problem of finding a sequence u_n such that $u_0 = 0$, $u_1 = 1$, and

$$u_{n+2} - 5u_{n+1} + 6u_n = 0$$

for all $n \geq 0$. Observe that we must have

$$u_{n+2}t^{n+2} - 5u_{n+1}t^{n+2} + 6u_nt^{n+2} = 0$$

for all $n \geq 0$, so that summing over all $n \geq 0$ yields

$$((\mathcal{G}u)(t) - u_1t - u_0) - 5(t(\mathcal{G}u)(t) - u_0) + 6t^2(\mathcal{G}u)(t) = 0.$$

Recalling that $u_0 = 0$, $u_1 = 1$, and rearranging, we obtain

$$(6t^2 - 5t + 1)(\mathcal{G}u)(t) = t.$$

Thus, using partial fractions, we obtain

$$\begin{aligned} (\mathcal{G}u)(t) &= \frac{t}{6t^2 - 5t + 1} = \frac{t}{(1-2t)(1-3t)} \\ &= \frac{-1}{1-2t} + \frac{1}{1-3t} \\ &= -\sum_{r=0}^{\infty} (2t)^r + \sum_{r=0}^{\infty} (3t)^r \\ &= \sum_{r=0}^{\infty} (3^r - 2^r)t^r. \end{aligned}$$

It follows that $u_r = 3^r - 2^r$.

Next consider the rather trivial problem of finding a sequence u_n such that $u_0 = 1$ and

$$(n+1)u_{n+1} + u_n = 0$$

for all $n \geq 0$. For every t we have

$$(n+1)u_{n+1}t^n + u_nt^n = 0$$

and so, summing over all n and assuming that the usual laws of differentiation apply to infinite sums, we obtain

$$(\mathcal{G}u)'(t) + (\mathcal{G}u)(t) = 0.$$

This differential equation gives $(\mathcal{G}u)(t) = Ae^{-t}$ for some constant A . Setting $t = 0$, we obtain

$$1 = u_0 = (\mathcal{G}u)(0) = Ae^0 = A.$$

Thus

$$(\mathcal{G}u)(t) = e^{-t} = \sum_{r=0}^{\infty} \frac{(-1)^r}{r!} t^r,$$

so $u_r = (-1)^r / r!$.

We can write down some of the correspondences between sequences and their \mathcal{G} -transforms:

$$\begin{aligned} (a_0, a_1, a_2, \dots) &\longleftrightarrow (\mathcal{G}a)(t), \\ (a_0 + b_0, a_1 + b_1, a_2 + b_2, \dots) &\longleftrightarrow (\mathcal{G}a)(t) + (\mathcal{G}b)(t), \\ a * b &\longleftrightarrow (\mathcal{G}a)(t)(\mathcal{G}b)(t), \\ (0, a_0, a_1, a_2, \dots) &\longleftrightarrow t(\mathcal{G}a)(t), \\ (a_1, 2a_2, 3a_3, \dots) &\longleftrightarrow (\mathcal{G}a)'(t). \end{aligned}$$

It is also important that we can recover the sequence a from its \mathcal{G} -transform. One way of seeing this is to note that

$$a_r = \frac{(\mathcal{G}a)^{(r)}(0)}{r!}.$$

We can use these rules, as in the examples above, to convert problems about sequences into problems about functions and vice versa. In textbooks and examinations, the effect of such a transformation is to make things simpler. In real life, it will usually convert your problem into a more complicated problem. However, occasionally you strike lucky and it is these occasions that make transforms such a valuable weapon in the mathematician's armory.

Up to now we have handled \mathcal{G} -transforms formally. However, if we wish to use the methods of analysis, we need to know that $\sum_{r=0}^{\infty} a_r t^r$ converges, at least when $|t|$ is small. Provided that the a_r do not increase too rapidly, this will always be the case. However, we run into difficulties when we try to extend our ideas to “two-sided sequences” (a_r) , where r runs through all integers rather than just the positive ones, and to the resulting sums $\sum_{r=-\infty}^{\infty} a_r t^r$. If $|t|$ is small, then $|t^r|$ is large when r is large and negative, while if $|t|$ is large, then $|t^r|$ is large when r is large and positive. In many cases, the best we can hope for is that $\sum_{r=0}^{\infty} a_r t^r$ might converge when $t = -1$ and $t = 1$. It is not very useful to talk about functions defined at only two points, but we save the situation by moving from \mathbb{R} to \mathbb{C} .

If we have a well-behaved sequence (a_r) of complex numbers where r runs through all integers, then we consider the sum $\sum_{r=-\infty}^{\infty} a_r z^r$, where the complex number z belongs to the unit circle (or, in other words, $|z| = 1$). Since any such z can be written

$$z = e^{i\theta} = \cos \theta + i \sin \theta$$

with $\theta \in \mathbb{R}$, it is more usual to talk about the 2π -periodic function $\sum_{r=-\infty}^{\infty} a_r e^{ir\theta}$. We thus have the

“Fourier series transform” (once again, the name is nonstandard) \mathcal{H} given by

$$(\mathcal{H}\mathbf{a})(\theta) = \sum_{r=-\infty}^{\infty} a_r e^{ir\theta}.$$

The \mathcal{H} -transform takes a two-sided sequence \mathbf{a} to a 2π -periodic complex-valued function $f = \mathcal{H}\mathbf{a}$ on the real line, but historically mathematicians have been more interested in reversing the process and obtaining \mathbf{a} from f . If

$$f(\theta) = \sum_{r=-\infty}^{\infty} a_r e^{ir\theta},$$

then, arguing formally,

$$\begin{aligned} \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\theta) e^{-ik\theta} d\theta &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{r=-\infty}^{\infty} a_r e^{i(r-k)\theta} d\theta \\ &= \sum_{r=-\infty}^{\infty} \frac{a_r}{2\pi} \int_{-\pi}^{\pi} e^{i(r-k)\theta} d\theta \\ &= \sum_{r=-\infty}^{\infty} \frac{a_r}{2\pi} \int_{-\pi}^{\pi} \cos(r-k)\theta + i \sin(r-k)\theta d\theta = a_k. \end{aligned}$$

If we write

$$\hat{f}(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\theta) e^{-ik\theta} d\theta,$$

then we obtain the celebrated Fourier sum formula

$$f(\theta) = \sum_{r=-\infty}^{\infty} \hat{f}(r) e^{ir\theta}. \quad (1)$$

DIRICHLET [VI.36] proved that this formula holds in its natural interpretation for reasonably well-behaved functions, but the question of the appropriate interpretation and proof for wider classes of functions took much longer to settle (see CARLESON'S THEOREM [V.5]). Aspects of the question are still open today.

It is worth noting that we can obtain qualitative information about a sequence from its \mathcal{H} -transform and vice versa without explicit calculation. For example, if $a_r r^{m+3}$ forms a bounded sequence, then the rules for term-by-term differentiation show that $\mathcal{H}\mathbf{a}$ is continuously m times differentiable, and if f is m times continuously differentiable, then repeated integration by parts shows that the numbers $r^m \hat{f}(r)$ form a bounded sequence.

Suppose that f represents a signal fed into a “black box,” such as a telephone system, which gives rise to a resultant signal Tf . Many important black boxes in physics and engineering have the “infinite linearity” property that

$$T\left(\sum_{r=-\infty}^{\infty} c_r g_r\right)(\theta) = \sum_{r=-\infty}^{\infty} c_r Tg_r(\theta)$$

for all well-behaved function g_r and constants c_r . Many such systems also have the key property that

$$Te_k(\theta) = \gamma_k e_k(\theta)$$

for some constant γ_k , where we have written $e_k(\theta)$ for the quantity $e^{-ik\theta}$. In other words, the functions e_k are EIGENFUNCTIONS [I.3 §4.3] for T . We can use the Fourier sum formula to obtain the formula

$$\begin{aligned} Tf(\theta) &= \left(\sum_{r=-\infty}^{\infty} \hat{f}(r) Te_r \right)(\theta) \\ &= \sum_{r=-\infty}^{\infty} \gamma_r \hat{f}(r) e_r(\theta). \end{aligned}$$

In this context, it makes sense to think of f as the weighted sum of simple signals e_k of frequency k .

Mathematicians are always interested to see what happens if sums are replaced by integrals. In this case we obtain the classical Fourier transform. If F is a reasonably well-behaved function $F: \mathbb{R} \rightarrow \mathbb{C}$, then we define its *Fourier transform* $\mathcal{F}F$ by the formula

$$\mathcal{F}F(\lambda) = \int_{-\infty}^{\infty} F(t) e^{-i\lambda t} dt.$$

Much of the analysis that is typically taught in the first year or two of a university mathematics course was developed in the context of this transform and related topics. Using that analysis, it is not hard to obtain the correspondences

$$\begin{aligned} F(t) &\longleftrightarrow (\mathcal{F}F)(\lambda), \\ F(t) + G(t) &\longleftrightarrow (\mathcal{F}F)(\lambda) + (\mathcal{F}G)(\lambda), \\ F * G(t) &\longleftrightarrow (\mathcal{F}F)(\lambda) (\mathcal{F}G)(\lambda), \\ F(t + u) &\longleftrightarrow e^{-iu\lambda} (\mathcal{F}F)(\lambda), \\ F'(t) &\longleftrightarrow i\lambda \mathcal{F}F(\lambda). \end{aligned}$$

In this context we define the convolution of F and G by

$$F * G(t) = \int_{-\infty}^{\infty} F(t-s) G(s) ds.$$

There is an element of truth in the saying that the importance of the Fourier transform is that it converts convolution to multiplication and the importance of convolution is that it is the operation that is transformed to multiplication by the Fourier transform. Just as we can use the G -transform to solve difference equations, we can use the \mathcal{F} -transform to solve important classes of PARTIAL DIFFERENTIAL EQUATIONS [I.3 §5.4] that occur in physics and some parts of probability theory. For more on the Fourier transform, see [III.27].

By rescaling the Fourier sum formula (1), we obtain the formula

$$F(t) = \sum_{r=-\infty}^{\infty} \frac{1}{2\pi N} \int_{-\pi N}^{\pi N} F(s) e^{-irs/N} ds e^{irt/N}$$

when $|t| < \pi N$. If we let $N \rightarrow \infty$, we obtain, more or less formally,

$$F(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} (\mathcal{F}F)(s) e^{ist} ds,$$

which translates to the marvelous formula

$$(\mathcal{F}\mathcal{F}F)(t) = 2\pi F(-t).$$

Like the Fourier sum formula, this *Fourier inversion formula* can be proved under a wide range of circumstances, though often at the price of reinterpreting the formula in novel ways.

Beautiful though the Fourier inversion formula is, it should be noted that, both in practice and in theory, we often need only the observation that $\mathcal{F}F = \mathcal{F}G$ implies $F = G$. The *uniqueness of the Fourier transform* is often easier to prove and more convenient to use, and it holds over a wider range of conditions than the inversion formula. A similar observation holds for other transforms.

When we talked about the Fourier sums associated with 2π -periodic functions, we said that $\hat{f}(r)$ measured the proportion of the signal f with frequency $2\pi r$. In the same way, $(\mathcal{F}F)(\lambda)$ gives a measure of the proportion of F composed of frequencies close to λ . There is a family of inequalities, known generically as *Heisenberg uncertainty principles*, which say, in effect, that if most of $\mathcal{F}F$ is concentrated in a narrow band, then the signal F must be very spread out. This fact places strong restrictions on our ability to manipulate signals and occupies a central place in quantum theory.

At the beginning of this article we talked about transformations of sequences and saw that it was easier to handle one-sided sequences than two-sided sequences. In the same way, we can apply Fourier transforms to a wider range of functions $F: \mathbb{R} \rightarrow \mathbb{C}$ if we know that $F(t) = 0$ for $t < 0$. More specifically, if F is such a one-sided function, and if it does not grow too fast, then we can compute the *Laplace transform*

$$\begin{aligned} (\mathcal{L}F)(x + iy) &= \int_{-\infty}^{\infty} F(s) e^{-(x+iy)s} ds \\ &= \int_0^{\infty} F(s) e^{-(x+iy)s} ds \end{aligned}$$

whenever x and y are real and x is sufficiently large. If we use the more natural notation

$$(\mathcal{L}F)(z) = \int_0^{\infty} F(s) e^{-zs} ds,$$

we see that $\mathcal{L}F$ can be considered as a weighted average of HOLOMORPHIC [I.3 §5.6] (that is, complex differentiable) functions and this can be used to show that $\mathcal{L}F$ is holomorphic. The Laplace transform shares

many of the properties of the Fourier transform and we can use these, as well as the extensive collection of results on holomorphic functions, whenever we manipulate Laplace transforms. Many of the deepest results in number theory, such as the PRIME NUMBER THEOREM [V.33], are most easily obtained by clever uses of the Laplace transform.

The transforms we have discussed all belong to the same family, as is indicated by the fact that they all take convolution to multiplication. The general idea of a transform has been developed in several different directions, generally by concentrating on some aspects of the “classical transforms” and being prepared to lose others.

One of the most important of these new transforms is the *Gelfand transform*, which gives a concrete representation of the abstract *commutative Banach algebras*. It is discussed in OPERATOR ALGEBRAS [IV.19 §3.1]. Other *integral transforms* extend the integral definition of the Fourier transform by setting up a correspondence

$$F(t) \longleftrightarrow \int_{-\infty}^{\infty} F(s) K(\lambda - s) ds$$

or, more generally,

$$F(t) \longleftrightarrow \int_{-\infty}^{\infty} F(s) \kappa(s, \lambda) ds.$$

Another interesting transform is the *Radon* or *x-ray transform*. We shall consider the three-dimensional case and talk very informally. Suppose we shine a beam of radiation through a body in direction \mathbf{u} . Suppose also that f is a function defined on \mathbb{R}^3 that represents how much radiation is absorbed by different parts of the body. What we can measure is the amount of radiation absorbed along any given straight line. We might present some of this information in the form of a two-dimensional image, which could represent the amount absorbed by all lines in the direction \mathbf{u} . In general, we can use f to define a new function

$$(\mathcal{R}f)(\mathbf{u}, \mathbf{v}) = \int_{-\infty}^{\infty} f(t\mathbf{u} + \mathbf{v}) dt,$$

which tells us how much radiation is absorbed along the line in direction \mathbf{u} that goes through a vector \mathbf{v} perpendicular to \mathbf{u} . The *tomography problem* deals with the recovery of f from $\mathcal{R}f$.

Because the idea of a transform has been developed in so many different directions, any attempt to give a general definition results in something too general to be useful. The most that we can say about the various transforms is that they present a more

or less distant analogy to the classical Fourier transforms and that this analogy has been found useful by those who developed them. (See also THE FOURIER TRANSFORM [III.27], SPHERICAL HARMONICS [III.89], REPRESENTATION THEORY [IV.12 §3], and WAVELETS AND APPLICATIONS [VII.3].)

III.94 Trigonometric Functions

Ben Green

The basic trigonometric functions “sin” and “cos,” as well as the four related functions “tan,” “cot,” “sec,” and “cosec,” will probably be familiar to most readers in some form. One way to define the sine function $\sin : \mathbb{R} \rightarrow [-1, 1]$ is as follows.

In almost all branches of mathematics one measures angles using *radians*, which are defined in terms of arc-length: to say that the angle $\angle AOB$ in figure 1 is θ radians is to say that the arc AB of the circle has length θ . This definition makes sense when $0 \leq \theta < 2\pi$. One then defines $\sin \theta$ to be the length PB, where P is the foot of the perpendicular from B to OA. It is very important that this length be taken with the correct *sign*. If $0 < \theta < \pi$ then we take the positive sign, whereas if $\pi < \theta < 2\pi$ we take the negative sign. In other words, $\sin \theta$ is the y -coordinate of the point B.

The sine function is, at the moment, defined on the interval $[0, 2\pi)$. To define it on all of \mathbb{R} one simply insists that it be periodic with period 2π (that is, that it satisfies the relation $\sin \theta = \sin(2\pi n + \theta)$ for any integer n).

There is one problem with our definition of sine. What do we mean by the *length* of the arc AB? The only really satisfactory way of understanding this is to use calculus. The equation of the unit circle is $y = \sqrt{1 - x^2}$, at least if (x, y) lies in the upper-right quadrant. (Otherwise one must be careful about sign.) The formula for the arc-length of a curve $y = f(x)$ between $y = a$ and $y = b$ is

$$S = \int_a^b \sqrt{1 + (dx/dy)^2} dy.$$

(This may be thought of as a definition, though the motivation for the definition comes from pictures.) For the circle, $\sqrt{1 + (dy/dx)^2} = 1/\sqrt{1 - y^2}$. Since the arc-length of the circle between the points P = $(x, \sin \theta)$ and A = $(1, 0)$ is θ , this gives the formula

$$\int_0^{\sin \theta} \frac{dy}{\sqrt{1 - y^2}} = \theta \quad (1)$$

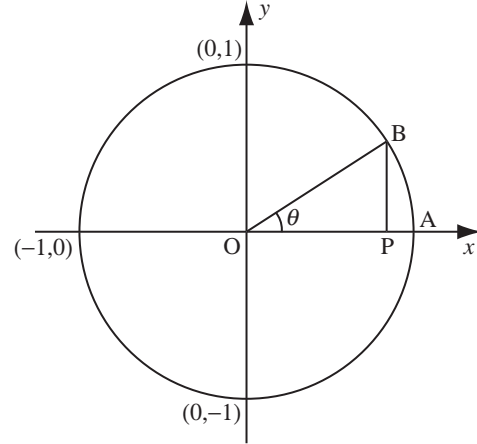


Figure 1 Interpreting trigonometric functions geometrically.

for $0 \leq \theta \leq \pi/2$ (we do not care about what x is). This can be regarded as giving a precise, even if implicit, definition of $\sin \theta$ for $0 \leq \theta < \pi/2$.

As with many of the most natural concepts in mathematics, \sin may be defined in a multitude of equivalent ways. Another definition (whose equivalence to the first one is not obvious) is

$$\sin z = z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \cdots \quad (2)$$

This infinite series converges for all real z . The resulting definition has a distinct advantage over (1), in that it also makes sense when z is an arbitrary *complex* number (that is why we replaced the letter θ by z). It therefore allows us to extend \sin to a HOLOMORPHIC FUNCTION [I.3 §5.6] on \mathbb{C} .

If the sine function is analytic, then what is its derivative? The answer is the cosine function $\cos z$, which may be defined in much the same way as \sin : either geometrically or using a power series. The power series is

$$\cos(z) = \frac{z^2}{2!} - \frac{z^4}{4!} + \frac{z^6}{6!} - \cdots, \quad (3)$$

which may be obtained by differentiating the series for \sin term by term (naturally, this is an operation that must be properly justified, but it can be).

If one differentiates again, one gets the formula $(d^2/dz^2) \sin z = -\sin z$. In fact, it is possible to define $\sin : \mathbb{R} \rightarrow [-1, 1]$ as the unique solution y to the differential equation $y'' = -y$ that also satisfies the initial value conditions $y(0) = 0$, $y'(0) = 1$. This is a very sensible way of proving that the two definitions (1) and

(2) are equivalent (it is a good calculus exercise to prove that $\sin'' = -\sin$ using (1)).

Ultimately, the power series expansions (2) and (3) display the most important side of \sin and \cos , which is their relation with the EXPONENTIAL FUNCTION [III.25]:

$$e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots.$$

Comparing this with (2) and (3), one gets the famous formula

$$e^{i\theta} = \cos \theta + i \sin \theta.$$

The exponential functions $\theta \mapsto e^{in\theta}$ are *characters*, that is, HOMOMORPHISMS [I.3 §4.1] from $\mathbb{R}/2\pi\mathbb{Z}$ to the unit circle S^1 (which form groups under addition mod 2π and multiplication, respectively). This makes them the natural objects with which to do a FOURIER ANALYSIS [III.27] of 2π -periodic functions on \mathbb{R} . Because \sin and \cos are real-valued, it is convenient to try to decompose such a function $f(x)$ not into exponentials, but as a series

$$a_0 + a_1 \cos x + b_1 \sin x + a_2 \cos 2x + b_2 \sin 2x + \cdots.$$

Under favorable circumstances (if the function f is sufficiently smooth, say) one can recover the coefficients a_i, b_i by using *orthogonality relations* such as

$$\frac{1}{\pi} \int_0^{2\pi} \cos nx \cos mx \, dx = \begin{cases} 0 & \text{for all } n, m \geq 0, \, n \neq m, \\ 1 & n = m, \end{cases}$$

and

$$\frac{1}{\pi} \int_0^{2\pi} \cos nx \sin mx \, dx = 0 \quad \text{for all } n, m \geq 0.$$

Thus, for example, we have

$$a_n = \frac{1}{\pi} \int_0^{2\pi} f(x) \cos nx \, dx.$$

Such decompositions into trigonometric functions ultimately underlie devices like compact disk players and mobile phones.

Let us conclude by remarking that there is a whole zoo of formulas concerning \sin , \cos , and the other four trigonometric functions (which we have not discussed here), as well as integrals involving these functions. It is these formulas that make the trigonometric functions an indispensable tool in classical Euclidean geometry. There are many further formulas in that setting. To mention just one beautiful example, the area of a triangle inscribed in a unit circle with angles A, B , and C is exactly $2 \sin A \sin B \sin C$.

Uncountable Sets

See COUNTABLE AND UNCOUNTABLE SETS [III.11]

III.95 Universal Covers

Let X be a TOPOLOGICAL SPACE [III.92]. A *loop* in X can be defined as a continuous function f from the closed interval $[0, 1]$ to X such that $f(0) = f(1)$. A *continuous family* of loops is a continuous function F from $[0, 1]^2$ to X such that $F(t, 0) = F(t, 1)$ for every t ; the idea is that for each t we can define a loop f_t by taking $f_t(s)$ to be $F(t, s)$, and if we do this then the loops f_t “vary continuously” with t . A loop f is *contractible* if it can be continuously shrunk to a point: more formally, there should be a continuous family of loops $F(t, s)$ with $F(0, s) = f(s)$ for every s and with all values of $F(1, s)$ equal. If all loops are contractible, then X is said to be *simply connected*. For instance, a sphere is simply connected, but a torus is not because there are loops that “go around” the torus and therefore cannot be contracted (since any continuous deformation of a loop that goes around the torus goes around the same number of times).

Given any path-connected space (that is, a space X such that any two points in X are linked by a continuous path), we can define a closely related *simply connected* space \tilde{X} as follows. First, we pick an arbitrary “base point” x_0 in X . We then take the set of all continuous paths f from $[0, 1]$ to X such that $f(0) = x_0$ (but we do not necessarily ask for $f(1)$ to be x_0). Next, we regard two of these paths f and g as *equivalent*, or *homotopic*, if $f(1) = g(1)$ and there is a continuous family of paths that begins with f and ends with g and always has the same beginning point and endpoint. That is, f and g are homotopic if there is a continuous function F from $[0, 1]^2$ to X such that $F(t, 0) = x_0$ and $F(t, 1) = f(1) = g(1)$ for every t , and $F(0, s) = f(s)$ and $F(1, s) = g(s)$ for every s . Finally, we define the *universal cover* \tilde{X} of X to be the space of all homotopy classes of paths: that is, it is the QUOTIENT [I.3 §3.3] of the space of all continuous paths that start at x_0 by the EQUIVALENCE RELATION [I.2 §2.3] of homotopy.

Let us see how this works in practice. As mentioned earlier, the torus is not simply connected, so what is its universal cover? To answer this question, it helps to think of the torus in a slightly artificial way: fix a point x_0 and define the torus to be the set of all continuous

PUP: whole article added after first proof of this part sent. Please check carefully.

PUP: Tim thinks that the context will indeed be clear to the reader.

PUP: I can confirm that \cos and then \sin is OK here.

paths that begin at x_0 , with two of these paths regarded as equivalent if they have the same endpoint. If we do this, then for each path, “all we care about” is where it ends, and the set of endpoints is clearly the torus itself. But this was not the definition of the universal cover. There we cared not just about the endpoint of a path but also about *how we reached* the endpoint. For instance, if the path happens to be a loop, in which case the endpoint is x_0 itself, then we care about how many times that loop goes around the torus, and in what manner it goes around.

The torus can be defined as the quotient of \mathbb{R}^2 by the equivalence relation where we define two points as equivalent if their difference belongs to \mathbb{Z}^2 . Then any point in \mathbb{R}^2 maps to a point in the torus (by the quotient map). Any continuous path on the torus then “lifts” uniquely to the plane in the following sense. Fix a point u_0 in \mathbb{R}^2 that maps to x_0 in the torus. Then if you trace out any continuous path in the torus that starts at x_0 , there will be exactly one way of tracing out a path in \mathbb{R}^2 such that each point in that path maps to the appropriate point in the path in the torus.

Now suppose that we have two paths in the torus that start at x_0 and end at the same point x_1 . Then the “lifts” of those paths both start at u_0 but all we know about their endpoints is that they are *equivalent*: we do not know that they are the same. Indeed, if the first path is a contractible loop and the second is a loop that goes once around the torus, then their lifts will end at *different* points. It turns out (and if you try to visualize this then you will see that the result is very natural and plausible) that the “lifts” of two paths will end at the same point if and only if the original paths are homotopic. In other words, there is a one-to-one correspondence between homotopy classes of paths in the torus and points in \mathbb{R}^2 . This shows that \mathbb{R}^2 is the universal cover of the torus. In a sense, the operation of passing from a space to its universal cover “unfolds” the quotienting operation that we use to get from the universal cover to the space.

As its name suggests, the universal cover has a universal property. Roughly speaking, a *cover* of a space X is a space Y and a continuous surjection from Y to X such that the inverse image of a small neighborhood in X is a disjoint union of small neighborhoods in Y . If U is the universal cover of X and Y is any other cover of X , then U can be made into a cover of Y in a natural way. For instance, one can define a cover of the torus by an infinite cylinder by wrapping the cylinder around, and the cylinder can in turn be covered by the plane.

An example of the use of universal covers can be found in GEOMETRIC AND COMBINATORIAL GROUP THEORY [IV.11 §§7, 8].

III.96 Variational Methods

Lawrence C. Evans

The *calculus of variations* is both a theory in itself and a toolbox of techniques for studying certain kinds of (often extremely nonlinear) ordinary and partial differential equations. These equations, which arise when we seek critical points of appropriate “energy” functionals, are usually far more tractable than other nonlinear problems.

1 Critical Points

Let us begin with a simple observation from first-year calculus, where we learn that if $f = f(t)$ is a smooth function defined on the real line \mathbb{R} and if f has a local minimum (or maximum) at a point t_0 , then $(df/dt)(t_0) = 0$.

The calculus of variations vastly extends this insight. The basic object to be considered is a *functional* F , which is applied not to real numbers but to functions, or rather to certain admissible classes of functions. That is, F takes functions u to real numbers $F(u)$. If u_0 is a minimizer of F (that is, $F(u_0) \leq F(u)$ for all admissible functions u), then we can expect that “the derivative of F at u_0 is zero.” Of course, this idea has to be made precise, which one might expect to be tricky since the space of admissible functions is infinite dimensional. But in practice these so-called variational methods end up using just standard calculus, and they provide deep insights into the nature of minimizing functions u_0 .

2 One-Dimensional Problems

The simplest situation to which variational techniques apply involves functions of a single variable. Let us see why minimizers of appropriate functionals in this setting must automatically satisfy certain ordinary differential equations.

2.1 Shortest Distance

As a warmup problem, we shall show that the shortest path between two points in the plane is a line segment. Of course, this is obvious, but the methods we develop can be applied to much more interesting situations.

Suppose, then, that we are given two points x and y in the plane. We take as our class of admissible functions all smooth, real-valued functions u , defined on some interval $I = [a, b]$, such that $u(a) = x$ and $u(b) = y$. The length of this curve is

$$F[u] = \int_I (1 + (u')^2)^{1/2} dx, \quad (1)$$

where $u = u(x)$ and a prime denotes differentiation with respect to x . Now suppose that some particular curve u_0 minimizes the length. We want to deduce that the graph of u_0 is a line segment, which we will do by “setting the derivative of F to zero” at the minimizer u_0 .

To make sense of this idea, select any other smooth function w that is defined on our interval I and that vanishes at its endpoints. For each t define $f(t)$ to be $F[u_0 + tw]$. Since the graph of the function $u_0 + tw$ connects the given endpoints, and since u_0 gives the minimum length, it follows that the function f , which is just an ordinary function from \mathbb{R} to \mathbb{R} , has a minimum at $t = 0$. Therefore, $(df/dt)(0) = 0$. But we can explicitly compute $(df/dt)(0)$ by differentiating under the integral sign and then integrating by parts. This gives

$$\int_I \frac{u'_0 w'}{(1 + (u'_0)^2)^{1/2}} dx = - \int_I \left(\frac{u'_0}{(1 + (u'_0)^2)^{1/2}} \right)' w dx.$$

This identity holds for all functions w with the properties specified above, and consequently

$$\left(\frac{u'_0}{(1 + (u'_0)^2)^{1/2}} \right)' = \frac{u''_0}{(1 + (u'_0)^2)^{3/2}} = 0 \quad (2)$$

everywhere on the interval I .

To summarize the discussion so far: if the graph of u_0 minimizes the distance between the given endpoints, then u''_0 identically equals zero, and therefore the shortest path is a line segment. This conclusion may not seem too exciting, but even this simple case has an interesting feature. The calculus of variations automatically focuses our attention on the expression

$$\kappa = \frac{u''}{(1 + (u')^2)^{3/2}},$$

which turns out to be the *curvature* of the graph of u . The graph of the minimizer u_0 has zero curvature everywhere.

2.2 Generalization: The Euler-Lagrange Equations

It turns out that the technique we used for the previous example is extremely powerful and can be vastly generalized.

One useful extension is to replace the length functional (1) by a more general functional of the form

$$F[u] = \int_I L(u', u, x) dx, \quad (3)$$

where $L = L(v, z, x)$ is a given function, sometimes called the *Lagrangian*. Then $F[u]$ can be interpreted as the “energy” (or “action”) of a given function u defined on the interval I .

Suppose next that a particular curve u_0 is a minimizer of F , subject to certain fixed boundary conditions. We want to extract information about the behavior of u_0 , and to do so we proceed as in our first example. We select a smooth function w as above, define $f(t) = F[u_0 + tw]$, observe that f has a minimum at $t = 0$, and consequently deduce that $(df/dt)(0) = 0$. As in the previous calculation, we then explicitly compute this derivative:

$$\frac{df}{dt}(0) = \int_I L_v w' + L_z w dx = \int_I (-(L_v)' + L_z) w dx.$$

Here, L_v and L_z stand for the partial derivatives $\partial L / \partial v$ and $\partial L / \partial z$, evaluated at (u'_0, u_0, x) . This expression equals zero for all functions w satisfying the given conditions. Therefore,

$$-(L_v(u'_0, u_0, x))' + L_z(u'_0, u_0, x) = 0 \quad (4)$$

everywhere on the interval I . This nonlinear ordinary differential equation for the function u_0 is called the *Euler-Lagrange equation*. The key point is that any minimizer of our functional F must be a solution of this differential equation, which often contains important geometrical or physical information.

For example, take $L(v, z, x) = \frac{1}{2}mv^2 - W(z)$, which we interpret as the difference between the kinetic energy and the potential energy W of a particle of mass m moving along the real line. The Euler-Lagrange equation (4) is then

$$mu''_0 = -W'(u_0),$$

which is *Newton's second law of motion*. The calculus of variations provides us with an elegant derivation of this fundamental law of physics.

2.3 Systems

We can generalize further, by taking

$$F[\mathbf{u}] = \int_I L(\mathbf{u}', \mathbf{u}, x) dx, \quad (5)$$

where now we are taking vector-valued functions \mathbf{u} that map the interval I into \mathbb{R}^m . If \mathbf{u}_0 is a minimizer in

some appropriate class of functions, then one can compute the Euler-Lagrange equation using ideas similar to those discussed above. We obtain the equations

$$-(L_{v^k}(\mathbf{u}'_0, \mathbf{u}_0, x))' + L_{z^k}(\mathbf{u}'_0, \mathbf{u}_0, x) = 0, \quad (6)$$

one for each k . Here L_{v^k} and L_{z^k} represent the partial derivatives of L with respect to the k th variables of \mathbf{u}' and \mathbf{u} , evaluated at $(\mathbf{u}'_0, \mathbf{u}_0, x)$. These equations form a system of coupled ordinary differential equations for the components of $\mathbf{u}_0 = (u_0^1, \dots, u_0^m)$.

For a geometric example, put

$$L(v, z, x) = \left(\sum_{i,j=1}^m g_{ij}(z) v^i v^j \right)^{1/2},$$

so that $F[\mathbf{u}]$ is the length of the curve \mathbf{u} in the RIEMANNIAN METRIC [I.3 §6.10] determined by the g_{ij} . When \mathbf{u}_0 is a curve of constant unit speed, the Euler-Lagrange system of equations (6) can be rewritten, after some work, to read

$$(u_0^k)'' + \sum_{i,j=1}^m \Gamma_{ij}^k(u_0^i)'(u_0^j)' = 0 \quad (k = 1, \dots, m)$$

for certain expressions Γ_{ij}^k , called *Christoffel symbols*, that can be computed in terms of the g_{ij} . Solutions of this system of ordinary differential equations are called *geodesics*. Thus, we have deduced that *length-minimizing curves are geodesics*.

A physical example is $L(v, z, x) = \frac{1}{2}m|v|^2 - W(z)$, for which the Euler-Lagrange equation is

$$m\mathbf{u}_0'' = -\nabla W(\mathbf{u}_0).$$

This is Newton's second law of motion for a particle in \mathbb{R}^m moving under the influence of the potential energy W .

3 Higher-Dimensional Problems

Variational methods also apply to expressions involving functions of several variables, in which case the resulting Euler-Lagrange equations are *partial* differential equations (PDEs).

3.1 Least Area

A first example extends our earlier examination of shortest curves. For this problem we are given a region U in the plane, with boundary ∂U , and a real-valued function g defined on the boundary. We then look at a class of admissible real-valued functions u , defined on U , with the condition that u should equal g on the

boundary. We can think of the graph of u as a two-dimensional curved surface with a boundary equal to the graph of g . The area of this surface is

$$F[u] = \int_U (1 + |\nabla u|^2)^{1/2} dx. \quad (7)$$

Let us assume that a particular function u_0 minimizes the area among all other surfaces with the given boundary. What can we deduce about the geometric behavior of this so-called *minimal surface*?

Yet again we proceed by writing $f(t) = F[u_0 + tw]$, differentiating with respect to t , and so on. After some calculation we eventually discover that

$$\operatorname{div} \left(\frac{\nabla u_0}{(1 + |\nabla u_0|^2)^{1/2}} \right) = 0 \quad (8)$$

within the region U , where “div” denotes the divergence operator. This nonlinear PDE is the *minimal surface equation*. The left-hand side turns out to be a formula for (twice) the *mean curvature* of the graph of u_0 . Consequently, we have shown that *a minimal surface has zero mean curvature everywhere*.

Minimal surfaces are sometimes regarded physically as the surfaces formed by soap films when they are stretched between a fixed wire frame that traces out the boundary specified by the function g .

3.2 Generalization: The Euler-Lagrange Equations

It is now straightforward, and sometimes very profitable, to replace the area functional (7) by the general expression

$$F[u] = \int_U L(\nabla u, u, x) dx, \quad (9)$$

in which we now take U to be a region in \mathbb{R}^n . Assuming that u_0 is a minimizer, subject to given boundary conditions, we deduce the *Euler-Lagrange equation*

$$-\operatorname{div}(\nabla_v L(\nabla u_0, u_0, x)) + L_z(\nabla u_0, u_0, x) = 0. \quad (10)$$

This is a nonlinear PDE that a minimizer must satisfy. A given PDE is called *variational* if it has this form.

If, for example, we take $L(v, z, x) = \frac{1}{2}|v|^2 + G(z)$, the corresponding Euler-Lagrange equation is the *nonlinear Poisson equation*

$$\Delta u = g(u),$$

where $g = G'$ and $\Delta u = \sum_{k=1}^n u_{x_k x_k}$ is the LAPLACIAN [I.3 §5.4] of u . We have shown that this important PDE is variational. This is a valuable insight, since we can then find solutions by constructing minimizers (or other critical points) of the functional $F[u] = \int_U \frac{1}{2}|\nabla u|^2 + G(u) dx$.

4 Further Issues in the Calculus of Variations

Our examples have shown pretty convincingly how useful our simple method, called computing the *first variation*, can be when applied to the right geometrical and physical problems. And indeed, variational principles and methods appear in several branches of mathematics and physics. Many of the objects that mathematicians consider most important have an underlying variational principle of some kind. The list is impressive and, besides the examples we have discussed, includes Hamilton's equations, the Yang-Mills and Selberg-Witten equations, various nonlinear wave equations, Gibbs states in statistical physics, and dynamic programming equations from optimal control theory.

Many issues remain. For example, if $f = f(t)$ has a local minimum at a point t_0 , then we know not only that $(df/dt)(t_0) = 0$, but also that $(d^2f/dt^2)(t_0) \geq 0$. The attentive reader will correctly guess that a generalization of this observation, called computing the *second variation*, is important for the calculus of variations. It provides an insight into appropriate convexity conditions that are needed to ensure that critical points are in fact stable minimizers. Even more fundamental is the question of the existence of minimizers or other critical points. Here mathematicians have devoted great ingenuity to designing appropriate function spaces within which "generalized" solutions can be found. But these weak solutions need not be smooth, and so the further question of their regularity and/or possible singularities must then be addressed.

However, these are all highly technical mathematical issues, far beyond the scope of this article. We end our discussion here, in the hope that our excessive demands upon the reader's attention have been minimized.

III.97 Varieties

Two simple examples of varieties are the circle and the parabola, which can be defined by the polynomial equations $x^2 + y^2 = 1$ and $y = x^2$, respectively. With one qualification, a variety is the solution set of a system of polynomial equations. The qualification is that there are certain examples that we do not want to include. For instance, the set of solutions to the equation $x^2 - y^2 = 0$ is the union of the two lines $x = y$ and $x = -y$, which naturally splits into two pieces. So the solution set to a system of polynomial equations

is called an *algebraic set*, and it is called a *variety* if it cannot be written as a union of smaller algebraic sets.

The examples just given were subsets of the plane \mathbb{R}^2 . However, the concept is much more general: varieties can live in \mathbb{R}^n for any n , and also in \mathbb{C}^n for any n . Indeed, the definitions make sense, and are interesting and important, in \mathbb{F}^n , where \mathbb{F} can be any field.

The varieties defined so far have been *affine* varieties. For many purposes it is more convenient to deal with *projective* varieties. The definition is similar, but now they live inside a PROJECTIVE SPACE [III.74], and the polynomials used to define them must be homogeneous—that is, any multiple of a solution must still be a solution.

See ALGEBRAIC GEOMETRY [IV.7] and ARITHMETIC GEOMETRY [IV.6] for more information.

III.98 Vector Bundles

Let X be a TOPOLOGICAL SPACE [III.92]. A vector bundle over X is, roughly speaking, a way of associating a vector space with each point x of X in such a way that these spaces "vary continuously" as you vary x . As an example, consider a smooth surface X in \mathbb{R}^3 . Associated with each point x is the *tangent plane* at x , which varies continuously with x and can be identified in a natural way with a two-dimensional vector space. A more precise definition is as follows: a *vector bundle of rank n* over X is a topological space E , together with a continuous map $p : E \rightarrow X$, such that the inverse image $p^{-1}(x)$ of each point x (that is, the set of points in E that map to x) is an n -dimensional vector space. The most obvious vector bundle of rank n over X is the space $\mathbb{R}^n \times X$ with the map $p(v, x) = x$; this is called the *trivial bundle*. However, the interesting bundles are the nontrivial ones, such as the tangent bundle of the 2-sphere. One can learn a great deal about a topological space by understanding its vector bundles. For this reason, vector bundles are central to algebraic topology. See ALGEBRAIC TOPOLOGY [IV.10 §5] for more details.

III.99 Von Neumann Algebras

A *unitary representation* of a GROUP [I.3 §2.1] G is a HOMOMORPHISM [I.3 §4.1] that associates with each element g of G a UNITARY MAP [III.52 §3.1] U_g defined on some HILBERT SPACE [III.37] H . A von Neumann algebra is a special kind of C^* -ALGEBRA [III.12], intimately connected with the theory of unitary representations.

There are several equivalent ways of defining von Neumann algebras. One is as follows. It can be checked that, given any unitary representation, its *commutant*, defined to be the set of all OPERATORS [III.52] in $B(H)$ that commute with every single unitary map U_g in the representation, forms a C^* -algebra. Von Neumann algebras are algebras that arise in this way. They can also be defined abstractly as follows: a C^* -algebra A is a von Neumann algebra if there is a BANACH SPACE [III.64] X such that the DUAL [III.19 §4] of X is A (when A is itself considered as a Banach space).

The basic building blocks of von Neumann algebras are special kinds of von Neumann algebras called *factors*. The classification of factors is a major topic of research, which includes some of the most celebrated theorems of the second half of the twentieth century. See OPERATOR ALGEBRAS [IV.19 §2] for more details.

III.100 Wavelets

If you wish to send a black and white picture from one computer to another, then an obvious way of doing it is to encode it pixel by pixel: 0 for black and 1 for white. However, for certain pictures this would obviously be extremely inefficient. For instance, if the picture is a square, the left half of which is entirely white and the right half of which is entirely black, then it is clearly much better to send a set of instructions for reconstructing the picture than a list of every single pixel. Furthermore, the precise details of the pixels usually do not matter: if you want a patch of gray, then it is enough to put in black and white pixels in the right proportion and make sure that they are evenly distributed.

However, finding a good way of encoding pictures is difficult, and an important area of research in engineering. A picture can be thought of as a function from a rectangle to \mathbb{R} . The set of all such functions forms a VECTOR SPACE [I.3 §2.3], and a natural way to try to come up with a good encoding is to find a good basis for this space. Here “good” means that the functions one is interested in (that is, ones that correspond to the kinds of pictorial representations one might wish to send) are determined by just a few of their coefficients, apart from minor variations that are not detectable by the human eye.

Wavelets are a particularly good basis for many purposes. In some ways they are like FOURIER TRANSFORMS [III.27], but they are much better suited to encoding details such as sharp boundaries, and patterns that are “localized,” rather than spread throughout the picture.

For more details, see WAVELETS AND APPLICATIONS [VII.3].

III.101 The Zermelo–Fraenkel Axioms

The Zermelo–Fraenkel, or ZF, axioms are a collection of axioms that provide a foundation for set theory. They may be viewed in two ways. The first is as a list of the “allowed operations” on sets. For example, there is an axiom that states that, given sets x and y , there exists a “pair set,” whose members are precisely x and y .

One of the reasons the ZF axioms are important is that it is possible to reduce all of mathematics to set theory, so the ZF axioms can be regarded as a foundation for mathematics as a whole. Of course, for this to be the case it is vital that the operations allowed by the ZF axioms do indeed allow one to perform all of the usual mathematical constructions. Some of the axioms are rather subtle as a result.

The other way to view the ZF axioms is as giving us just what we need to “build up” the world of all sets, starting with just the empty set. One can look at the various ZF axioms and see that each one plays an essential role as we create the set-theoretic universe. Equivalently, they are “closure rules” that any universe of sets, or more precisely any model of set theory, should obey. So, for example, there is an axiom that says that every set has a power set (the set of all its subsets), and this axiom allows us to build up a huge collection of sets starting with just the empty set: one obtains the power set of the empty set, the power set of the power set of the empty set, and so on. Indeed, the universe of all sets could (in a sense) be described as the closure of the empty set under all the allowable operations of ZF.

The ZF axioms are written in the language of FIRST-ORDER LOGIC [IV.2 §1]. So each axiom can mention variables (which are interpreted as ranging over all sets), as well as the usual logical operations, and also one “primitive relation,” namely membership. For example, the pair-set axiom above would be formally written as

$$(\forall x)(\forall y)(\exists z)(\forall t)(t \in z \iff t = x \text{ or } t = y).$$

By convention, the ZF axioms do not include the AXIOM OF CHOICE [III.1]; when one does include the axiom of choice, the axioms are usually called the “ZFC axioms.”

For a more detailed discussion of the ZF axioms see SET THEORY [IV.1 §3.1].

Part IV

Branches of Mathematics

IV.1 Algebraic Numbers

Barry Mazur

The roots of our subject go back to ancient Greece while its branches touch almost all aspects of contemporary mathematics. In 1801 the *Disquisitiones Arithmeticae* of CARL FRIEDRICH GAUSS [VI.26] was first published, a “founding treatise,” if ever there was one, for the modern attitude toward number theory. Many of the still unachieved aims of current research can be seen, at least in embryonic form, as arising from Gauss’s work.

This article is meant to serve as a companion to the reader who might be interested in learning, and thinking about, some of the classical theory of algebraic numbers. Much can be understood, and much of the beauty of algebraic numbers can be appreciated, with a minimum of theoretical background. I recommend that readers who wish to begin this journey carry in their backpacks Gauss’s *Disquisitiones Arithmeticae* as well as Davenport’s *The Higher Arithmetic* (1992), which is one of the gems of exposition of the subject, and which explains the founding ideas clearly and in depth using hardly anything more than high-school mathematics.

1 The Square Root of 2

The study of algebraic numbers and algebraic integers begins with, and constantly reverts back to, the study of ordinary rational numbers and ordinary integers. The first algebraic irrationalities occurred not so much as *numbers* but rather as *obstructions* to simple answers to questions in geometry.

That the ratio of the diagonal of a square to the length of its side cannot be expressed as a ratio of whole numbers is purported to be one of the vexing discoveries of the early Pythagoreans. But this very ratio, when squared, is 2:1. So we might—and later mathematicians certainly did—deal with it algebraically. We can think of this ratio as a cipher, about which we know nothing

beyond the fact that its square is 2 (a viewpoint taken toward algebraic numbers by KRONECKER [VI.48], as we shall see below). We can write $\sqrt{2}$ in various forms, e.g.,

$$\sqrt{2} = |1 - i|, \quad (1)$$

and we can think of $1 - i = 1 - e^{2\pi i/4}$ as the world’s simplest trigonometric sum; we shall see generalizations of this for all quadratic surds below. We can also view $\sqrt{2}$ as a limit of various infinite sequences, one of which is given by the elegant CONTINUED FRACTION [III.22]

$$\sqrt{2} = 1 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2 + \dots}}} \quad (2)$$

Directly connected to this continued fraction (2) is the Diophantine equation

$$2X^2 - Y^2 = \pm 1 \quad (3)$$

known as the *Pell equation*. There are infinitely many pairs of integers (x, y) satisfying this equation, and the corresponding fractions y/x are precisely what you get by truncating the expression in (2). For example, the first few solutions are $(1, 1)$, $(2, 3)$, $(5, 7)$, and $(12, 17)$, and

$$\left. \begin{aligned} \frac{3}{2} &= 1 + \frac{1}{2} = 1.5, \\ \frac{7}{5} &= 1 + \frac{1}{2 + \frac{1}{2}} = 1.4, \\ \frac{17}{12} &= 1 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2}}} = 1.416\dots \end{aligned} \right\} \quad (4)$$

Replace the ± 1 on the right-hand side of (3) by *zero* and you get $2X^2 - Y^2 = 0$, an equation all of whose positive real-number solutions (X, Y) have the ratio $Y/X = \sqrt{2}$, so it is easy to see that the sequence of fractions (4) (these being alternately larger and smaller than $\sqrt{2} = 1.414\dots$) converges to $\sqrt{2}$ in the limit. Even more striking is that (4) is a list of fractions that best approximate $\sqrt{2}$. (A rational number a/d is said to be a *best approximant* to a real number α if a/d is closer to α than any rational number of denominator smaller than or equal to d .) To deepen the pic-

of ± 1 when divided by 5, and *minus* otherwise.

What governs the choice of the plus terms and minus terms is whether or not n is a *quadratic residue modulo* 5. Here is a brief explanation of this terminology. If m is an integer, two integers a, b are said to be *congruent modulo* m (in symbols we write $a \equiv b \pmod{m}$) if the difference $a - b$ is an integral multiple of m ; if a, b , and m are positive numbers, it is equivalent to ask that a and b have the same “remainder” (sometimes also called “residue”) when each is divided by m (see MODULAR ARITHMETIC [III.60]). An integer a relatively prime to m is called a *quadratic residue modulo* m if a is congruent to the square of some integer, modulo m ; otherwise it is called a *quadratic nonresidue modulo* m . So, 1, 4, 6, 9, ... are quadratic residues modulo 5, while 2, 3, 7, 8, ... are quadratic nonresidues modulo 5.

A generalization of equations (5) and (10) (the “analytic formula for the L -function attached to quadratic Dirichlet characters”) gives a very surprising formula for the conditionally convergent sum of terms $\pm 1/n$, where n runs through positive integers relatively prime to a fixed integer and the sign of $\pm 1/n$ corresponds to whether n is a quadratic residue, or nonresidue modulo that integer.

3 Quadratic Irrationalities

The quadratic formula

$$X = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

gives the solutions (usually two) to the general quadratic polynomial equation $aX^2 + bX + c = 0$ as a rational expression of the number \sqrt{D} , where $D = b^2 - 4ac$ is known as the *discriminant* of the polynomial $aX^2 + bX + c$, or, equivalently, of the corresponding homogeneous QUADRATIC FORM [III.75] $aX^2 + bXY + cY^2$. This formula introduces many irrational numbers: Plato’s dialogue “Theaetetus” has the young Theaetetus credited with the discovery that \sqrt{D} is irrational whenever D is a natural number that is not a perfect square. The curious switch, from initially perceiving an *obstruction* to a problem to eventually embodying this obstruction as a *number* or an *algebraic object of some sort* that we can effectively study, is repeated over and over again, in different contexts, throughout mathematics. Much later, *complex* quadratic irrationalities also made their appearance. Again these were not at first regarded as “numbers as such,” but rather as *obstructions* to the solution of problems. Nicholas Chuquet, for example,

in his 1484 manuscript, *Le Triparty*, raised the question of whether or not there is a number whose triple is four plus its square and he comes to the conclusion that there is no such number because the quadratic formula applied to this problem yields “impossible” numbers, i.e., complex quadratic irrationalities in our terminology.²

For any real quadratic (“integral”) irrationality there is a discussion along similar lines to the ones we have just given (expressions (1)–(5) for $\sqrt{2}$ and expressions (6)–(10) for $\frac{1}{2}(1 + \sqrt{5})$). For complex irrationalities, there is also such a theory, but with interesting twists. For one thing, we do not have anything directly comparable to continued-fraction expansions for a complex quadratic irrationality. In fact, the simple, but true, answer to the problem of how to find an infinite number of rational numbers that converge to such an irrationality is that you cannot! Correspondingly, the analogue of the Pell equation has only finitely many solutions. As a consolation, however, the appropriate “analytic formula” has a simpler sum, as we will see below.

Let d be any square-free integer, positive or negative. Associated with d is a particularly important number τ_d , defined as follows. If d is congruent to 1 mod 4 (that is, if $d - 1$ is a multiple of 4), then $\tau_d = \frac{1}{2}(1 + \sqrt{d})$; otherwise, $\tau_d = \sqrt{d}$. We will refer to these quadratic irrationalities τ_d as *fundamental algebraic integers of degree* 2. The general notion of an “algebraic integer” is defined in section 11. An algebraic integer of degree two is simply a root of a quadratic polynomial of the form $X^2 + aX + b$ with a, b ordinary integers. In the first case (when $d \equiv 1$ modulo 4), τ_d is a root of the polynomial $X^2 - X + \frac{1}{4}(1 - d)$ and in the second it is a root of $X^2 - d$. The reason special names are given to these quadratic irrationalities is that *any* quadratic algebraic integer is a linear combination (with ordinary integers as coefficients) of 1 and one of these fundamental quadratic algebraic integers.

4 Rings and Fields

I think that one of the big early advances in mathematics is the now-current, universal recognition of the importance of studying the properties of *collections* of mathematical objects, and not just the objects in isolation. A *ring* R of complex numbers is a collection of

2. BOMBELLI [VI.8], in the sixteenth century, would refer to irrational square roots, of positive or of negative numbers, as “deaf” (reminiscent of the word *surd* that is still in use) and as “numbers impossible to name.”

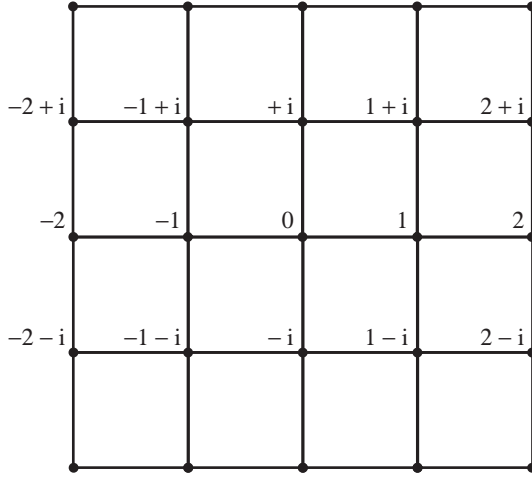


Figure 2 The Gaussian integers are the vertices of this lattice of squares tiling the complex plane.

them that contains 1 and is closed under the operations of addition, subtraction, and multiplication. That is, if a, b are any two numbers in R , $a \pm b$ and ab must also be in R . If such a ring R has the further property that it is closed under division by nonzero elements (i.e., if a/b is again in R whenever a and b are, and $b \neq 0$), then we say that R is a *field*. (These concepts are discussed further in *FIELDS* [I.3 §2.2] and *RINGS, IDEALS, AND MODULES* [III.83].) The ring \mathbb{Z} of ordinary integers, $\{0, \pm 1, \pm 2, \dots\}$ is our “founding example” of a ring; visibly, it is the smallest ring of complex numbers.

The collection of all real or complex numbers that are integral linear combinations of 1 and τ_d is closed under addition, subtraction, and multiplication, and is therefore a ring, which we denote by R_d . That is, R_d is the set of all numbers of the form $a + b\tau_d$ where a and b are ordinary integers. These rings R_d are our first, basic, examples of *rings of algebraic integers* beyond that prototype, \mathbb{Z} , and they are the most important rings that are receptacles for quadratic irrationalities. Every quadratic irrational algebraic integer is contained in exactly one R_d .

For example, when $d = -1$ the corresponding ring R_{-1} , usually referred to as the ring of *Gaussian integers*, consists of the set of complex numbers whose real and imaginary parts are ordinary integers. These complex numbers may be visualized as the vertices of the infinite tiling of the complex plane by squares whose sides have length 1 (see figure 2).

When $d = -3$ the complex numbers in the corresponding ring R_{-3} may be visualized as the vertices of

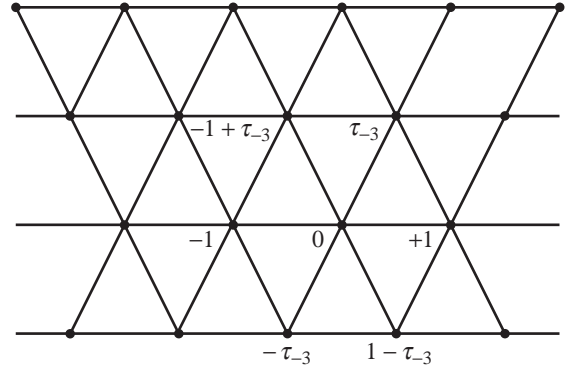


Figure 3 The elements of the ring R_{-3} are the vertices of this lattice of hexagons tiling the complex plane.

the regular hexagonal tiling of the complex plane (see figure 3).

With the rings R_d in hand, we may ask ring-theoretic questions about them, and here is some of the standard vocabulary useful for this. A *unit* u in a given ring R of complex numbers is a number in R whose reciprocal $1/u$ is also in R ; a *prime* (or synonymously, an *irreducible*) element in R is a nonunit that cannot be written as the product of two nonunits in R . A ring of complex numbers R has the *unique factorization property* if every nonzero, nonunit, algebraic number in R can be expressed as a product of prime elements in exactly one way (where two factorizations are counted as the same if one can be obtained from the other by rearranging the order in which the primes appear and multiplying them by units).

In the prototype ring \mathbb{Z} of ordinary integers, the only units are ± 1 . The fundamental fact that any ordinary integer greater than 1 can be uniquely expressed as a product of (positive) prime numbers (that is, that \mathbb{Z} enjoys the unique factorization property) is crucial for much of the number theory done with ordinary integers. That this unique factorization property for integers actually required proof was itself a hard-won realization of Gauss, who also provided its proof (see *THE FUNDAMENTAL THEOREM OF ARITHMETIC* [V.16]).

It is easy to see that there are only four units in the ring R_{-1} of Gaussian integers, namely ± 1 and $\pm i$; multiplication by any of these units effects a *symmetry* of the infinite square tiling (figure 2 above). There are only six units in the ring R_{-3} , namely ± 1 , $\pm \frac{1}{2}(1 + \sqrt{-3})$ and $\pm \frac{1}{2}(1 - \sqrt{-3})$; multiplication by any of these units results in a symmetry of the infinite hexagonal tiling (figure 3 above).

Fundamental to understanding the arithmetic of R_d is the following question: which ordinary prime numbers p remain prime in R_d and which ones factorize into products of primes in R_d ? We will see shortly that if a prime number does factorize in R_d , it must be expressible as the product of precisely two prime factors. For example, in the ring of Gaussian integers, R_{-1} , we have the factorizations

$$\begin{aligned} 2 &= (1 + i)(1 - i), \\ 5 &= (1 + 2i)(1 - 2i), \\ 13 &= (2 + 3i)(2 - 3i), \\ 17 &= (1 + 4i)(1 - 4i), \\ 29 &= (2 + 5i)(2 - 5i), \\ &\vdots \end{aligned}$$

where all the Gaussian integer factors in brackets above are *prime* in the ring of Gaussian integers.

Let us say that an odd prime p *splits* in R_{-1} if it factorizes into a product of at least two primes and *remains prime* if it does not do so. As we shall soon see, the officially agreed-upon definitions of splitting and remaining prime for more general rings of algebraic integers (even ones of the form R_d) are worded slightly, but very significantly, differently from the way we have just defined these concepts in the ring R_{-1} of Gaussian integers. (Note that we have excluded the prime $p = 2$ from the above dichotomy. This is because 2 *ramifies* in R_{-1} ; for a discussion of this concept see section 7 below.) In any event, there is an elementary computable *rule* that tells us, for any R_d , which primes p split and which remain prime in this agreed sense. The rule depends upon the residue of p modulo $4d$: the reader is invited to guess it for the ring of Gaussian integers given the data just displayed above. In general, an elementary computable rule that says which primes split and which do not in a ring of algebraic integers such as R_d is referred to as a *splitting law* for the ring of algebraic integers in question.

5 The Rings R_d of Quadratic Integers

There is a very important “symmetry,” or AUTOMORPHISM [I.3 §4.1], defined on the ring R_d . It sends \sqrt{d} to $-\sqrt{d}$, keeps all ordinary integers fixed, and more generally, for rational numbers u and v , it sends $\alpha = u + v\sqrt{d}$ to what we may call its *algebraic conjugate* $\alpha' = u - v\sqrt{d}$. (The word “algebraic” is to remind you that this is not necessarily the same as the complex-conjugate symmetry of the complex numbers!)

You can immediately work out the formulas for this algebraic conjugation operation on the fundamental quadratic irrationalities τ_d : if d is not congruent to 1 modulo 4, then $\tau_d = \sqrt{d}$, so obviously $\tau'_d = -\tau_d$, while if d is congruent to 1 modulo 4, then $\tau_d = \frac{1}{2}(1 + \sqrt{d})$ and $\tau'_d = \frac{1}{2}(1 - \sqrt{d}) = 1 - \tau_d$. This symmetry $\alpha \mapsto \alpha'$ respects all algebraic formulas. For example, to work out the algebraic conjugate of a polynomial expression like $\alpha\beta + 2\gamma^2$, where α , β , and γ are numbers in R_d , you just replace each individual number by its algebraic conjugate, obtaining the expression $\alpha'\beta' + 2\gamma'^2$.

The most telling integer quantity attached to a number $\alpha = x + y\tau_d$ in R_d is its *norm* $N(\alpha)$, which is defined to be the product $\alpha\alpha'$. This equals $x^2 - d\gamma^2$ when $\tau_d = \sqrt{d}$ and $x^2 + xy - \frac{1}{4}(d-1)\gamma^2$ when $\tau_d = \frac{1}{2}(1 + \sqrt{d})$. The norm turns out to be *multiplicative*, meaning that $N(\alpha\beta) = N(\alpha)N(\beta)$, as you can directly check by multiplying out the formula for the norm of each factor and comparing with the norm of the product. This gives us a useful tactic for trying to factorize algebraic numbers in R_d , and offers criteria for determining whether a number α in R_d is a unit, and whether it is prime in R_d . In fact, an element $\alpha \in R_d$ is a unit if and only if $N(\alpha) = \alpha\alpha' = \pm 1$; in other words, the units are given by the integral solutions to the equations

$$X^2 - dY^2 = \pm 1 \quad (11)$$

or

$$X^2 + XY - \frac{1}{4}(d-1)Y^2 = \pm 1 \quad (12)$$

following the two cases. Here is the proof of this. If $\alpha = x + y\tau_d$ is a unit in R_d , then its reciprocal, $\beta = 1/\alpha$, must also be in R_d , and, of course, we have $\alpha\beta = 1$. Applying the norm to both sides of this equation and using the multiplicative property discussed above, we see that $N(\alpha)$ and $N(\beta)$ are reciprocal ordinary integers. Therefore, they are either both equal to +1 or both equal to -1. This shows that (x, y) is a solution to whichever of equation (11) or (12) is appropriate. In the other direction, if $N(\alpha) = \alpha\alpha' = \pm 1$, then the reciprocal of α is simply $\pm\alpha'$. This is in R_d so α is indeed a unit in R_d .

These homogeneous quadratic forms, the left-hand sides of equations (11) and (12) (which generalize formulas (3) and (9)), play an important role; let us refer to whichever of them is relevant to R_d as the *fundamental quadratic form* for R_d , and to its discriminant D as the *fundamental discriminant*. (D is equal to d if d is congruent to 1 modulo 4 and to $4d$ otherwise.) When d is negative there are only finitely many units (if $d < -3$ the only ones are ± 1) but when d is positive,

PUP: again we'd like to keep 'brackets' here, instead of changing to 'parentheses'. OK?

so that R_d consists entirely of real numbers, there are infinitely many. The ones that are greater than 1 are powers of a smallest such unit, ε_d , and this is called the *fundamental unit*.

For example, when $d = 2$ the fundamental unit, ε_2 , is $1 + \sqrt{2}$, and when $d = 5$ it is the golden mean, $\varepsilon_5 = \frac{1}{2}(1 + \sqrt{5})$. Since any power of a unit is again a unit, we immediately have a machine for producing infinitely many units from any single one. For example, taking powers of the golden mean, we get

$$\begin{aligned}\varepsilon_5 &= \frac{1}{2}(1 + \sqrt{5}), & \varepsilon_5^2 &= \frac{1}{2}(3 + \sqrt{5}), \\ \varepsilon_5^3 &= 2 + \sqrt{5}, & \varepsilon_5^4 &= \frac{1}{2}(7 + 3\sqrt{5}), \\ \varepsilon_5^5 &= \frac{1}{2}(11 + 5\sqrt{5}),\end{aligned}$$

all of which are units in R_5 . The study of these fundamental units was already under way in the twelfth century in India, but in general their detailed behavior as d varies still holds mysteries for us today. For example, there is a deep theorem of Hua (1942) that tells us that $\varepsilon_d < (4e^2d)^{\sqrt{d}}$ (for a proof of it along with a historical discussion of such estimates, see chapters 3 and 8 in Narkiewicz (1973)). There are examples of d that come close to attaining that bound, but we still do not know whether or not there is a positive number η and an infinity of square-free d for which $\varepsilon_d > d^{\eta}$. (The answer to this question would be yes if, for example, there were an infinity of R_d satisfying the unique factorization property! This follows from a famous theorem of Brauer (1947) and Siegel (1935); for a proof of the Brauer–Siegel theorem, see theorem 8.2 of chapter 8 in Narkiewicz (1973) or Lang (1970).)

6 Binary Quadratic Forms and the Unique Factorization Property

The principle of unique factorization is an all-important fact for the ring of ordinary integers \mathbb{Z} . The question of whether this principle does or does not hold for a given ring R_d is central to the algebraic number theory. There are helpful, analyzable, *obstructions* to the validity of unique factorization in R_d . These obstructions, in turn, connect with profound arithmetic issues, and have become the focus of important study in their own right. One such mode of expressing the obstruction to unique factorization is already prominent in Gauss’s *Disquisitiones Arithmeticae* (1801), in which much of the basic theory of R_d was already laid down.

This “obstruction” has to do with how many “essentially different” binary quadratic forms $aX^2 + bXY +$

cY^2 there are with discriminant equal to the fundamental discriminant D of R_d . (Recall that the discriminant of $aX^2 + bXY + cY^2$ is $b^2 - 4ac$, and that D equals $4d$ unless $d \equiv 1 \pmod{4}$, in which case it equals d .)

In order to define a binary quadratic form $aX^2 + bXY + cY^2$ of discriminant D , what you need to provide is simply a triplet of coefficients (a, b, c) such that $b^2 - 4ac = D$. Given such a form, one can use it to define other ones. For example, if we make a small linear change of the variables, replacing X by $X - Y$ and keeping Y fixed, then we get $a(X - Y)^2 + b(X - Y)Y + cY^2$, which simplifies to $aX^2 + (b - 2a)XY + (c - b + a)Y^2$. That is, we get a new binary quadratic form whose triplet of coefficients is $(a, b - 2a, c - b + a)$, and which (as can easily be checked) has the same discriminant D . We can “reverse” this change by replacing X by $X + Y$ and keeping Y fixed. If we do this reversal and perform the corresponding simplification then we get back our original binary quadratic form. Because of this reversibility, these two quadratic forms take exactly the same set of integer values as X and Y vary: it is therefore reasonable to think of them as *equivalent*.

More generally, then, one says that two binary quadratic forms are equivalent if one can be turned into the other (or minus the other) by any “reversible” linear change of variables with integer coefficients. That is, one chooses integers r, s, u, v such that $rv - su = \pm 1$, replaces X and Y by the linear combinations $X' = rX + sY$, $Y' = uX + vY$, and simplifies the resulting expression to get a new triplet of coefficients. The condition $rv - su = \pm 1$ guarantees that by a similar operation we can get back to our original binary quadratic form, and also that the new binary quadratic form has the same discriminant D as the old one. So when we talk of “essentially different” binary quadratic forms of discriminant D we mean that we cannot turn one into the other by this kind of change of variables.

Here is the surprising obstruction to unique factorization that Gauss discovered.

The unique factorization principle is valid in R_d if and only if every homogeneous quadratic form $aX^2 + bXY + cY^2$ with discriminant equal to the fundamental discriminant of R_d is equivalent to the fundamental quadratic form of R_d .

Furthermore, the collection of inequivalent quadratic forms whose discriminant is the fundamental discriminant of R_d expresses in concrete terms the degree to which R_d “enjoys unique factorization.”

If you have never seen this theory of binary quadratic forms before, try your hand at working with quadratic forms in the case where $D = -23$. The idea is to start with some particular quadratic form $aX^2 + bXY + cY^2$ of your choice with discriminant $D = b^2 - 4ac = -23$. Then, using a sequence of carefully chosen linear changes of variables you reduce the size of the coefficients a , b , and c until you can go no further. Eventually you should end up with one of the two (inequivalent) quadratic forms that there are with discriminant -23 : the fundamental form $X^2 + XY + 6Y^2$, or the form $2X^2 + XY + 3Y^2$. For example, can you see that the binary quadratic form $X^2 + 3XY + 8Y^2$ is equivalent to $X^2 + XY + 6Y^2$?

This type of exercise offers a small hint of the role that the *geometry of numbers* will play in the eventual theory. As you might expect from the venerability of these ideas, elegant streamlined methods have been discovered for making such calculations. Nevertheless, it is an open secret that any working mathematician, contemporary or ancient, engaged in this subject or nearby subjects, has done a myriad of straightforward simple hand computations along the lines of the above exercise.

If you try a few examples of this exercise, as I hope you do, here is one way of organizing your calculations. First, find a simple reversible linear change of variables to turn your form into an equivalent one with $a, b, c \geq 0$. (You may also have to multiply the whole form by -1 .)

The cleanest way of writing down all binary quadratic forms given by triplets (a, b, c) of discriminant -23 is to list the triplets in increasing order of b , which will now be an odd positive integer. For each value of b you can then choose a and c in such a way that their product is $\frac{1}{4}(b^2 + 23)$. At this point the aim is to build up a repertoire of moves that tend to decrease b (which will keep a and c within bounds as well). A big clue, and aid, here is that for any pair of relatively prime integers x, y if you evaluate your quadratic form $aX^2 + bXY + cY^2$ at $(X, Y) = (x, y)$ to get the integer $a' = ax^2 + bxy + cy^2$, you can find, for appropriate b' and c' , a quadratic form $a'X^2 + b'XY + c'Y^2$ equivalent to yours, with first coefficient a' . So, one tactic is to look for small integers represented by your quadratic form. Also the “example” linear change of variables $X \mapsto X - Y$, $Y \mapsto Y$ will lead you to be able to reduce the coefficient b to an integer smaller than $2a$. Can you check that $X^2 + XY + 6Y^2$ and $2X^2 + XY + 3Y^2$ are inequivalent?

Now, as we have just discussed, it follows from the general theory that R_{-23} does not have the unique factorization property. We can also see this directly. For example,

$$\tau_{-23} \cdot \tau'_{-23} = 2 \cdot 3,$$

and all four of the factors in this equation are irreducible in R_{-23} . To be a faithful companion, I should at this point give at least a hint at what connection there might be between this specific “failure of unique factorization” and the previous discussion. It may become a bit clearer in the next paragraph, but the underlying tension in the equation $\tau_{-23} \cdot \tau'_{-23} = 2 \cdot 3$ is that all the factors in our ring are prime: we are *missing* any elements in our ring R_{-23} that could factorize it further. We lack, for example, elements that play the role of the *greatest common divisor* of factors of this equation. The general theory regarding these matters (which we are not entering into here, but see EUCLID’S ALGORITHM [III.22]) tells us that what is missing is some element y in R_{-23} that is both a linear combination of the numbers τ_{-23} and 2 (with coefficients in the ring R_{-23}) and also a common divisor of τ_{-23} and 2 in the ring R_{-23} , i.e., such that τ_{-23}/y and $2/y$ are both in R_{-23} . There is no such element, for its norm must divide $N(\tau_{-23}) = 6$ and $N(2) = 4$, and therefore be equal to 2 , which can easily be shown to be impossible. But we are interested, rather, in the phenomenon that *inequivalence* of certain binary quadratic forms will indeed show this, so let us go on.

First, check that any linear combination

$$\alpha \cdot \tau_{-23} + \beta \cdot 2$$

with α, β elements of R_{-23} can also be written as $u \cdot \tau_{-23} + v \cdot 2$, where u and v are ordinary integers. Now compute the binary quadratic form given by systematically taking the norms of these linear combinations, and viewing these norms as functions of the integer coefficients u, v :

$$\begin{aligned} N(u \cdot \tau_{-23} + v \cdot 2) &= (\tau_{-23}u + 2v)(\tau'_{-23}u + 2v) \\ &= 6u^2 + 2uv + 4v^2. \end{aligned}$$

Viewing the u and the v as *variables*, and dubbing them U and V to emphasize their status as variables, we can say that the *norm quadratic form* obtained from the collection of linear combinations of τ_{-23} and 2 is

$$6U^2 + 2UV + 4V^2 = 2 \cdot (3U^2 + UV + 2V^2).$$

Now suppose that, contrary to fact, there *were* a common divisor, y , as above; in particular, the multiples of y in the ring R_{-23} would then be precisely the linear

combinations of the numbers τ_{-23} and 2. We would then have another way of describing those linear combinations; namely, for any pair of ordinary integers (u, v) there would be a pair of ordinary integers (r, s) such that

$$u \cdot \tau_{-23} + v \cdot 2 = y \cdot (r\tau_{-23} + s) = ry\tau_{-23} + sy.$$

Taking norms, as above, we would get

$$\begin{aligned} N(y \cdot (r\tau_{-23} + s)) &= N(ry\tau_{-23} + sy) \\ &= N(y)(6r^2 + rs + s^2). \end{aligned}$$

Again, thinking of r and s as variables and renaming them R and S we would have the corresponding norm quadratic form:

$$N(y) \cdot (6R^2 + RS + S^2) = 2 \cdot (6R^2 + RS + S^2).$$

Given the above facts—dependent, of course, on the contrary-to-fact hypothesis that there is a y as above—the key idea is that there would be linear changes of variables from (U, V) to (R, S) and back that would establish an equivalence between the two quadratic forms $2 \cdot (3U^2 + UV + 2V^2)$ and $2 \cdot (6R^2 + RS + S^2)$. But these quadratic forms are not equivalent! Their inequivalence therefore shows that the putative y does not exist and factorization in the ring R_{-23} is not unique.

7 Class Numbers and the Unique Factorization Property

In the previous section we saw that the collection of inequivalent quadratic forms of discriminant equal to the fundamental discriminant provides us with an obstruction to unique factorization. Somewhat later, a more articulated version of this obstruction arose, known as the *ideal class group* H_d of R_d . As its name implies, to describe this we must use the vocabulary of IDEALS [III.83 §2] and GROUPS [I.3 §2.1]. A subset I of R_d is an *ideal* if it has the following closure properties: if α belongs to I , so do $-\alpha$ and $\tau_d \alpha$, and if α and β belong to I , so does $\alpha + \beta$. (The first and third properties imply together that any integer combination of α and β belongs to I .) The basic example of such an ideal is the set of all multiples of some fixed, nonzero element y of R_d , where by a *multiple* of y we mean the product of y and an element of R_d . We denote this set tersely as (y) , or, slightly more expressively, as $y \cdot R_d$. An ideal of this sort, i.e., one that can be expressed as the set of all multiples of a single nonzero element y , is called a *principal ideal*. For example, the ring R_d itself is an ideal (it consists, after all, of all linear combinations of 1 and τ_d) and is even a principal ideal: in our

laconic terminology, it can be denoted $(1) = 1 \cdot R_d = R_d$. Strictly speaking, the singleton $\{0\}$ is also an ideal, but the ones that will interest us are the *nonzero ideals*.

As a direct counterpart to the obstruction principle involving binary quadratic forms that was described in the previous section, we have the following obstruction principle involving ideals.

The unique factorization principle is valid in R_d if and only if every ideal in R_d is principal.

Reflecting on this, you can get a sense of why the word “ideal” might have been chosen. Every principal ideal in R_d is of the form $y \cdot R_d$ for some number y in R_d (which is uniquely determined apart from multiplication by units), but sometimes there are more general ideals. These arise if you ever have two elements of R_d (think of τ_{-23} and 2, as in the previous section) such that the set of all their integer combinations *cannot* be expressed as the set of multiples of some fixed number y in R_d . This phenomenon is a sign that we may be missing numbers in R_d that provide fine enough factorizations to make the arithmetic in R_d as smooth going as one might hope for. Just as a principal ideal $y \cdot R_d$ corresponds to the number y , ideals of this more general kind (think of the set of all integer combinations of τ_{-23} and 2) can be thought of as corresponding to “ideal numbers” that should, “by rights,” be present in our ring, but happen not to be.

Once we think of ideals as standing for ideal numbers it makes some sense to try to multiply them: if I, J are two ideals in R_d , we let $I \cdot J$ denote the set of all finite sums of products $\alpha \cdot \beta$ in which α is in I and β is in J . The product of two principal ideals $(y_1) \cdot (y_2)$ is the principal ideal $(y_1 \cdot y_2)$ so, just as one would hope, multiplication of principal ideals corresponds to multiplication of the corresponding numbers. Multiplication of any ideal I by the ideal (1) leaves I unchanged: $(1) \cdot I = I$; we therefore refer to the ideal (1) as *the unit ideal*. With this new notion of *multiplication of ideals* we can now give the general definition of what it means for a prime number p to split or to remain prime in a ring R_d , the definition we promised in section 4.

The idea behind the definition is to use multiplication of ideals rather than of numbers. So if we are thinking about a prime p , the first thing we do is turn our attention to the principal ideal (p) in R_d . If this can be factorized as a product of two different ideals (*not necessarily principal ideals*, this is the whole point) in R_d , and if neither of these is the unit ideal $(1) = R_d$, then we say that p *splits* in R_d . If, on the other hand,

no factorization of the ideal (p) can be made without one of the factors being the ideal $(1) = R_d$, then we say that p *remains prime* in R_d . There is also a third important definition: if the principal ideal (p) can be expressed as the square of another ideal I , then we say that p *ramifies* in R_d . Continuing with the momentum of this definition, we may say that an ideal P is a *prime ideal* if P cannot be “factorized” as the product of two ideals neither of which is the unit ideal. This definition makes sense whether or not P is principal, so we are subtly shifting our attention from the multiplicative arithmetic of the numbers in R_d to the ideals.

By definition, two ideals are in the same *ideal class* if when you multiply each by an appropriate principal ideal you get the same ideal as a result. This is a natural EQUIVALENCE RELATION [I.2 §2.3] on ideals. It is also one that *respects products*, meaning that if I and J are two ideals, then the ideal class of their product $I \cdot J$ depends only on the ideal classes of I and J . (In other words, if I' is in the same ideal class as I and J' is in the same ideal class as J , then $I' \cdot J'$ is in the same ideal class as $I \cdot J$.) We can therefore say what we mean by *multiplication of ideal classes*: to multiply two classes, pick an ideal from each, multiply those, and take the ideal class of the resulting product. The set H_d of ideal classes of R_d , given this operation of multiplication, forms an Abelian group, in the sense that the multiplication law we have just defined is associative and commutative, and there are inverses. The identity element is the principal ideal R_d itself. This group H_d , the *ideal class group*, directly measures the extent to which the ideals of the ring R_d are principal: roughly speaking it is what you get if you take the multiplicative structure of all ideals and “divide out” by the principal ones.

As was mentioned in section 6, there is a close connection between ideal classes and binary quadratic forms. To begin to see this, take an ideal I of R_d and write it as the set of all integer combinations of two elements α, β of R_d . Then consider the norm function on the elements of I , that is,

$$\begin{aligned} N(x\alpha + y\beta) &= (x\alpha + y\beta)(x\alpha' + y\beta') \\ &= \alpha\alpha'x^2 + (\alpha\beta' + \alpha'\beta)xy + \beta\beta'y^2. \end{aligned}$$

This is a binary quadratic form in the variable coefficients x and y . If you start with a different choice of α, β that generate I you get a different form, but the two forms are scalar multiples of two forms with discriminant D that are equivalent to one another. Even better,

the equivalence class of these forms depends only on the ideal class of I .

It can be shown that there are only a finite number of distinct ideal classes of R_d ; that is, the ideal class group H_d is finite. The number of its elements is denoted h_d and called the *class number* of R_d . So, the obstruction to unique factorization of R_d is given by the nontriviality of the group H_d ; equivalently, unique factorization holds for R_d if and only if its class number is 1. But whether or not H_d is trivial, its detailed group-theoretic structure is profoundly related to the arithmetic of R_d .

The class number enters into the generalizations of formulas (5) and (10) of section 1; that is, the *analytic formulas* we alluded to in that section. These formulas represent just the beginning of one of the ongoing chapters of our subject, and form a bridge between the world of discrete arithmetical issues and that of calculus, infinite series, and volumes of spaces, all of which can be attacked by the methods of COMPLEX ANALYSIS [I.3 §5.6]. Here is a sample of them.

- (i) If $d > 0$ is a square-free integer and D is either d or $4d$ according to whether d is congruent to 1 modulo 4 or not, then

$$h_d \cdot \frac{\log \varepsilon_d}{\sqrt{D}} = \sum_{n \geq 0} \pm \frac{1}{n},$$

where the integers n run through those that are relatively prime to D and the signs \pm are chosen in a way that depends only on the residue class of n modulo D .

- (ii) If $d < 0$ we have a somewhat simpler formula: there is no fundamental unit ε_d in R_d to contend with, but when $d = -1$ or -3 , there are more roots of unity than merely ± 1 . If w_d denotes the number of roots of unity in R_d , then $w_{-1} = 4$, $w_{-3} = 6$ and otherwise $w_d = 2$, and then one has a formula of the following type:

$$\frac{h_d}{w_d \sqrt{D}} = \sum_{n \geq 0} \pm \frac{1}{n}.$$

As d tends to $-\infty$ the class number h_d tends to infinity.

We have effective lower bounds for the growth of h_d but these lower bounds are probably still far from the actual growth (cf. Goldfeld 1985). The effective lower bounds that are known are exceedingly weak. They follow, however, from beautiful work of Goldfeld, and of Gross and Zagier: for every real number $r < 1$

there is a computable constant $C(r)$ such that $h_d > C(r) \log |D|^r$. Here is a sample:

$$h_d > \frac{1}{55} \prod_{p|D} \left(1 - \frac{2\sqrt{p}}{p+1}\right) \cdot \log |D|$$

if $(D, 5077) = 1$.

It is a striking lacuna in our theory that, even today, nobody knows how to prove that there are infinitely many values of $d > 0$ for which R_d enjoys the unique factorization property—particularly since we expect that more than three quarters of them do! Our expectations are even more precise than that, thanks to Henri Cohen and Hendrik Lenstra, who make use of certain probabilistic expectations (now known as the *Cohen–Lenstra heuristics*) to conjecture that the density of positive fundamental discriminants of class number 1 among all positive fundamental discriminants is 0.75446....

8 The Elliptic Modular Function and the Unique Factorization Property

A different obstruction to unique factorization in R_d is available when d is negative. Now R_d may be thought of as a lattice in the complex plane (see figure 3), which makes a wonderful tool available for us: the classical *elliptic modular function* of KLEIN [VI.57],

$$j(z) = e^{-2\pi iz} + 744 + 196\,884 e^{2\pi iz} + 21\,493\,760 e^{4\pi iz} + 864\,299\,970 e^{6\pi iz} + \dots \quad (13)$$

This function, also colloquially referred to as the “ j -function,” converges for complex numbers $z = x + iy$ with $y > 0$. If $z = x + iy$ and $z' = x' + iy'$ are two such complex numbers, then $j(z) = j(z')$ if and only if the lattice generated by z and 1 in the complex plane is the same as the lattice generated by z' and 1 (or, equivalently, $z' = (az + b)/(cz + d)$, where a, b, c , and d are ordinary integers such that $ad - bc = 1$). We can paraphrase this by saying that the value $j(z)$ depends only on, and characterizes, the lattice generated by z and 1.

It turns out (by a theorem of Schneider) that if an algebraic number $\alpha = x + iy$ with $y > 0$ has the property that $j(\alpha)$ is also algebraic, then α is a (complex) quadratic irrationality; and the converse is also true. In particular, since $\alpha = \tau_d$ is such a complex quadratic irrationality when d is negative, the value, $j(\tau_d)$, of the j -function on τ_d is an algebraic number—in fact, an algebraic integer. This will be of some importance for

our story. First, since the ring R_d as situated in the complex plane is simply the lattice generated by τ_d and 1, it follows from the previous paragraph that this value $j(\tau_d)$ will be the same if we replace τ_d by *any* element α of R_d , as long as the lattice generated by α and 1 is the entire ring R_d . More importantly, $j(\tau_d)$ is an algebraic integer of degree roughly comparable with the class number of R_d . In particular, it is an ordinary integer if and only if the ring R_d has the unique factorization property. (This result is one of the great applications of a classical theory known as *complex multiplication*.) In brief, here is yet another answer to the question of when the unique factorization principle holds for R_d when d is negative: if $j(\tau_d)$ is an ordinary integer, the answer is *yes*; otherwise it is *no*.

The search for the full list of negative values of d for which R_d has the unique factorization property makes a marvelous tale: there are precisely nine values of d for which it occurs (see below), but for over two decades number theorists, while knowing these nine, could prove only that there were no more than ten. The history of how the *nonexistence* of a possible tenth value of d was established, and reestablished, is one of the thrilling chapters in our subject. K. Heegner, in an article published in 1934, provided what he claimed was a proof of the nonexistence of the possible *tenth value of d* . However, Heegner’s proof was framed in somewhat unfamiliar language and was not understood by the mathematicians of the time. His paper and his purported proof were largely forgotten until the late 1960s, when the nonexistence of the tenth field was established (to the mathematical community’s satisfaction) by Stark (1967) and independently, via a different method, by Baker (1971). It was only then that mathematicians took a second and closer look at Heegner’s original article and discovered that he had indeed proven exactly what he claimed. Moreover, his proof offered an elegant direct conceptual road to an understanding of the underlying issue.

Here are the nine values of d :

$$d = -1, -2, -3, -7, -11, -19, -43, -67, -163.$$

And here are the corresponding nine values of $j(\tau_d)$:

$$j(\tau_d) = 2^6 3^3, 2^6 5^3, 0, -3^3 5^3, -2^{15}, -2^{15} 3^3, -2^{18} 3^3 5^3, -2^{15} 3^3 5^3 11^3, -2^{18} 3^3 5^3 23^3 29^3.$$

PUP: I can confirm that the fact that the second number is greater than the first in the sequence is OK here.

As Stark once pointed out, if, for some of these values of d , you simply “plug” τ_d into the power series expansion for j , you get rather surprising formulas. For

IV.1. Algebraic Numbers

example, when $d = -163$, then

$$e^{-2\pi i \tau_d} = -e^{\pi \sqrt{163}}$$

is the first term of the power series for $j(\tau_{-163})$ (see formula (13)). Since $j(\tau_{-163}) = -2^{18}3^35^323^329^3$ and since all the terms $e^{2\pi n \tau_d}$ ($n > 0$) that appear in the power series for the j -function are relatively small, we find that $e^{\pi \sqrt{163}}$ is incredibly close to an integer. Indeed, it is $2^{18}3^35^323^329^3 + 744 + \dots$, which works out as $262\,537\,412\,640\,768\,744 - \epsilon$, where the error term ϵ is less than 7.5×10^{-13} .

9 Representations of Prime Numbers by Binary Quadratic Forms

More often than you might expect, it turns out to be possible to translate difficult and/or somewhat artificial problems about ordinary integers into natural and tractable problems about larger rings of algebraic integers. My favorite elementary example of this type is the theorem due to FERMAT [VI.12] that if a prime number p may be expressed as a sum of two squares, $p = a^2 + b^2$ with $0 < a \leq b$, then it has only one such expression. (For example, $1^2 + 10^2$ is the only way of expressing the prime number 101 as the sum of two squares.) Moreover, a prime number p can be expressed as a sum of two squares if and only if $p = 2$ or p is of the form $4k + 1$. (The “only if” part of this is easy to see: since any square is congruent either to 0 or to 1 mod 4, an odd integer that is a sum of two squares is necessarily congruent to 1 mod 4.) These statements about ordinary integers can be translated into basic statements about the ring of Gaussian integers. For if we write $a^2 + b^2 = (a + ib)(a - ib)$, with $i = \sqrt{-1}$, then we can view $a^2 + b^2$ as the norm of the (conjugate) elements $a \pm ib$ in the ring of Gaussian integers. So, if p is a prime number that admits an expression as a sum of squares, $p = a^2 + b^2$, it follows that each of the elements $a \pm ib$ has norm a prime integer. It is easy to deduce that p is itself a prime in the ring of Gaussian integers. Indeed, any factorization of $a \pm ib$ into a product of two Gaussian integers would have the property that the norms of the factors are ordinary integers which multiply out to be the prime p , and this severely limits their possibilities: one of them has to be a unit.

In other words, whenever $p = a^2 + b^2$, then

$$p = (a + ib)(a - ib)$$

is a factorization of the ordinary integer prime p into a product of two Gaussian integer primes. The uniqueness part of Fermat’s theorem then follows from (in

fact, it is readily seen to be equivalent to) the unique factorization property of the ring R_{-1} of Gaussian integers. That any prime number p of the form $4k + 1$ admits such an expression as a sum of two squares follows from the *splitting law* for primes p in the ring of Gaussian integers: an odd prime number p is a norm, and hence splits into the product of two distinct primes, in the ring of Gaussian integers if and only if p is congruent to 1 mod 4. This result is just the beginning of an immense chapter of arithmetic.

10 Splitting Laws and the Race between Residues and Nonresidues

The simple *splitting law* for ordinary prime integers p in the ring of Gaussian integers, which states that p splits if $p \equiv 1 \pmod{4}$ and not if $p \equiv -1 \pmod{4}$, invites us to ask how often each of these cases occurs (see figure 4). DIRICHLET [VI.36] proved a famous theorem that says that there are infinitely many primes in the arithmetic progression $c, m + c, 2m + c, \dots$ if the integers m and c are relatively prime. A more precise version of his result gives a clear asymptotic answer to the question we have just asked: as x goes to infinity, the ratio of the number of primes less than x that split to the number that do not tends to 1. (See ANALYTIC NUMBER THEORY [IV.2 §4] for a further discussion of Dirichlet’s theorem.)

For fun, one might ask a fussier question: which type of prime less than x is actually in greater abundance, the nonsplit primes or the split ones (see figure 4)? To put some perspective on this, let us widen our query: for q equal either to 4 or to an odd prime, let $A(x)$ be the number of primes $\ell < x$ that are quadratic residues modulo q and let $B(x)$ be the number of primes $\ell < x$ that are quadratic nonresidues modulo q . Let $D(x) = A(x) - B(x)$ be the difference; what does $D(x)$ look like?

For an absorbing account of the history and status of this problem, see the article “Prime number races” by Andrew Granville and Greg Martin in *American Mathematical Monthly*.

11 Algebraic Numbers and Algebraic Integers

Now that we have seen the algebraic integers $j(\tau_d)$ for negative values of d , and have touched on trigonometric sums, we have a few hints that, as with ordinary integers, the deep structure of these rings of quadratic integers may be better understood within a larger context

PUP: what do you think of this sentence? The full reference details are available so perhaps I should add this to the further reading of this article and reword here instead, but this is how the author would prefer this to be cited.

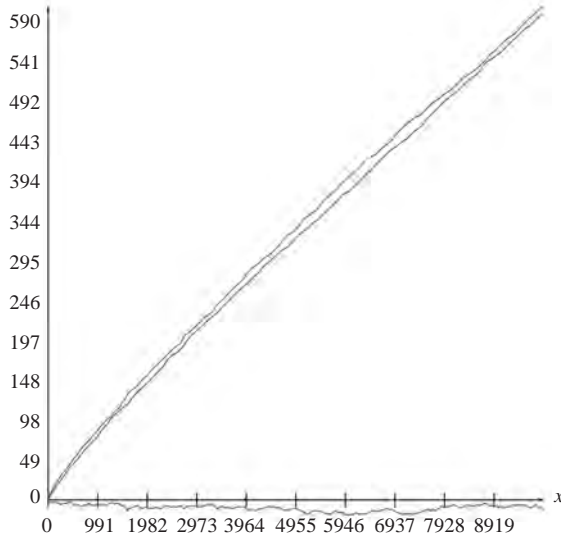


Figure 4 The higher of the two graphs in the figure represents the number of primes less than X that *remain prime* in the ring of Gaussian integers, and the lower represents the number of primes less than X that *split* in the ring of Gaussian integers. The third graph hovering around the x -axis represents the difference between the two numbers. We thank William Stein for this data.

of algebraic numbers. So now let us deal with algebraic numbers in full generality.

By a *monic* polynomial, we mean a polynomial of the form

$$P(X) = X^n + a_1X^{n-1} + \cdots + a_{n-1}X + a_n,$$

i.e., a polynomial of degree n such that the coefficient of X^n is 1. In general, the other coefficients are just assumed to be complex numbers. If $P(X) = X^n + a_1X^{n-1} + \cdots + a_{n-1}X + a_n$ is such a polynomial, and if θ is a complex number such that $P(\theta) = 0$, or, equivalently, if θ satisfies the polynomial equation

$$\theta^n + a_1\theta^{n-1} + \cdots + a_{n-1}\theta + a_n = 0,$$

we say that θ is a *root* of the polynomial $P(X)$. THE FUNDAMENTAL THEOREM OF ALGEBRA [V.15], initially proved by Gauss, guarantees that any such polynomial of degree n factors into a product of n linear polynomials. That is,

$$P(X) = (X - \theta_1)(X - \theta_2) \cdots (X - \theta_n)$$

for some complex numbers $\theta_1, \theta_2, \dots, \theta_n$ that are in fact precisely the roots of the polynomial $P(X)$.

If θ is a root of such a polynomial $P(X) = X^n + a_1X^{n-1} + \cdots + a_{n-1}X + a_n$ and if in addition the coeffi-

cients a_i are rational numbers, then θ is called an *algebraic number*. If the coefficients are not just rational but are in fact integers, then θ is called an *algebraic integer*. So, for example, the square root of any rational number is an algebraic number and the square root of any “ordinary” integer is an algebraic integer. The same holds true for n th roots of ordinary integers, or of algebraic integers, for any natural number n . For an example of a different sort, we have already mentioned the theorem that the values of the j -function on complex quadratic irrational integers are algebraic integers. For a (random) particular case of that theorem, the complex number $j(\tau_{-23})$ is a root of the monic polynomial

$$X^3 + 3\,491\,750X^2 - 5\,151\,296\,875X + 12\,771\,880\,859\,375.$$

An exercise: show that any algebraic number can be expressed as an algebraic integer divided by an ordinary integer.

12 Presentation of Algebraic Numbers

In dealing with any mathematical concept, we confront, in one way or another, the dual problem of the various forms in which it comes to us when it arises in our work, and the various ways we can *present* it so as to deal with it effectively. We have already seen a bit of this at the outset of this article, in our discussion of quadratic surds, and we will continue to see it in our treatment of them below, where the various modes in which quadratic surds can be presented—as *radicals*, as *eventually recurrent continued fractions*, or as *trigonometric sums*—come together, all contributing to their unified theory.

This issue of presentation is all the more of a problem with algebraic numbers in general, which may come to us in a multitude of ways. For example, they can arise as the coordinates of points on specific algebraic varieties whose defining equations may not be easily available, or as special values of functions like the j -function. It is natural, then, to look for some uniform way of presenting algebraic numbers, and the history of the subject shows how much effort has been devoted to such a search. For example, consider the focus on iterated radical expressions, as in the famous formula for the solution to the general cubic equation $X^3 = bX + c$ given by

$$X = \left(\frac{c}{2} + \sqrt{\frac{c^2}{2} - \frac{b^3}{27}}\right)^{1/3} + \left(\frac{c}{2} - \sqrt{\frac{c^2}{2} - \frac{b^3}{27}}\right)^{1/3}, \quad (14)$$

or the corresponding general solution to the fourth-degree equation. These were major achievements of sixteenth-century Italian algebra, and they culminated in the proof that the general fifth-degree algebraic number could *not* be so expressed, which was a major achievement of the early nineteenth century (see THE INSOLUBILITY OF THE QUINTIC [V.24]). The challenge to give *some* analytic expression for such fifth-degree algebraic numbers was the source of a classic book by Klein, *The Icosahedron*, written in the late nineteenth century. Kronecker wrote that it was the “dream of his youth” (his *Jugendtraum*) to establish a uniform mode of presentation for a class of algebraic numbers that interested him, by expressing them as values of certain analytic functions.

13 Roots of Unity

A central role in the theory of algebraic numbers is played by the *roots of unity*, that is, the n complex solutions of the equation $X^n = 1$, or equivalently the n roots of the polynomial $X^n - 1$. If we let $\zeta_n = e^{2\pi i/n}$, then these roots are precisely ζ_n and its powers, so in particular they are algebraic integers. They give us the factorization

$$X^n - 1 = (X - 1)(X - \zeta_n)(X - \zeta_n^2) \cdots (X - \zeta_n^{n-1}).$$

Now the powers of ζ_n form the vertices of a regular n -gon in the complex plane, centered at the origin. This has the following consequence, noticed by Gauss in his youth. It can be shown that compass and straight-edge constructions allow us, in effect, to extract square roots, so whenever ζ_n can be given as an expression built out of just square roots and the usual arithmetical operations, we have, implicitly, a ruler-and-compass construction of the regular n -gon, and conversely.

To get some idea of why square roots are so closely connected with these constructions, consider this. If we have given ourselves a *unit measure*, which we can view as the distance between the numbers 0 and 1 in the (complex) plane, and if we have already constructed, by whatever device, a specific point, x say, between 0 and 1 on the horizontal axis of the plane, we can first “construct” $x/2$ by straightedge and compass, and then go on to form a right-angled triangle with hypotenuse of length $1 + x/2$ and one of its other sides of length $1 - x/2$ (again using a straightedge and compass). The Pythagorean theorem gives us that the third side of that triangle is of length \sqrt{x} . If one follows this line of thought (but adapts it to deal with complex quantities

as well as the real number x as in the example we have just discussed), then one can see that the equations

$$\zeta_3 = \frac{1}{2}(1 + i\sqrt{3}),$$

$$\zeta_4 = \sqrt{i},$$

$$\zeta_5 = \frac{1}{4}(\sqrt{5} - 1) + i\frac{1}{8}(\sqrt{5 + \sqrt{5}}),$$

$$\zeta_6 = -\frac{1}{2}(1 + i\sqrt{3})$$

provide (implicit) constructions of the equilateral triangle, the square, the regular pentagon, and the regular hexagon, respectively. By contrast, ζ_7 cannot be expressed solely in terms of the arithmetical operations and square roots (it is the root of a quadratic equation with coefficients that are rational expressions in the roots of the irreducible *cubic* polynomial $X^3 - \frac{7}{3}X + \frac{7}{27}$), which already suggests that the regular heptagon might fail to be constructible by the standard classical means—and indeed it does fail without some act of “angle trisection.” (In principle, though, the reader can work out an expression for ζ_7 in terms of square roots and cube roots by means of the information provided in the parenthetical phrase above, together with equation (14).)

Gauss showed that if $n > 2$ is a prime number then the regular n -gon is classically constructible if and only if n is a *Fermat prime*, that is, a prime number of the form $2^{2^a} + 1$. So, for example, the 11-gon and 13-gon are not constructible by classical means, but since ζ_{17} is expressible as nested rational expressions of square roots, the 17-gon is, famously, constructible.

So, not all roots of unity can be expressed as iterated rational expressions of square roots. However, this inhospitability is not mutual, since all square roots of integers can be expressed as integer combinations of roots of unity. More mysteriously, the elusive fundamental units ε_d (for d positive), for which there is no known formula, are intimately related to a unit c_d in R_d which is an explicit rational expression of roots of unity. (See below: it is called a *circular unit*.) This satisfies the elegant formula

$$c_d = \varepsilon_d^{h_d}, \quad (15)$$

which establishes yet another explicit test of unique factorization: the equality $c_d = \varepsilon_d$ is a “litmus” requirement for the unique factorization principle to hold in R_d .

To give the flavor of the formulas involved, let p be an odd prime number and let a be an integer not divisible by p . Then define $\sigma_p(a)$ to be $+1$ if a is a *quadratic residue modulo p* , that is, if a is congruent to

the square of an integer modulo p , and -1 if not. The simple trigonometric sums of (1) and (6) generalize to *quadratic Gauss sums*:

$$\begin{aligned} \pm i^{(p-1)/2} \sqrt{p} = & \zeta_p + \sigma_p(2)\zeta_p^2 + \sigma_p(3)\zeta_p^3 + \cdots \\ & + \sigma_p(p-2)\zeta_p^{p-2} + \sigma_p(p-1)\zeta_p^{p-1}. \end{aligned} \quad (16)$$

This formula is not too hard to prove, apart from determining which sign is correct in the initial \pm , but after considerable efforts Gauss managed to work this out too. To see the connection between, say, formula (6) and (16) note that when $p = 5$, the left-hand side of (16) is $\sqrt{5}$ and the right-hand side is

$$\zeta_5 + -\zeta_5^2 - \zeta_5^{-2} + \zeta_5^{-1} = 2 \cos \frac{2}{5}\pi - 2 \cos \frac{4}{5}\pi.$$

As for the circular unit c_p , it is defined to be

$$\prod_{a=1}^{(p-1)/2} (\zeta_p^a - \zeta_p^{-a})^{\sigma_p(a)} = \prod_{a=1}^{(p-1)/2} \sin(\pi a/p)^{\sigma_p(a)},$$

and this leads to further formulas. For example, when $p = 5$, we have $\varepsilon_p = \tau_5 = \frac{1}{2}(1 + \sqrt{5})$, and since $h_5 = 1$, formula (6) for $p = 5$ tells us that

$$\frac{1 + \sqrt{5}}{2} = \frac{\zeta_5 - \zeta_5^{-1}}{\zeta_5^2 - \zeta_5^{-2}} = \frac{\sin \frac{1}{5}\pi}{\sin \frac{2}{5}\pi}.$$

14 The Degree of an Algebraic Number

If θ is an algebraic integer that is also a rational number, then θ is an “ordinary” integer. Here is the proof of this fact. If θ is a rational number, then we may write $\theta = C/D$ as a fraction in lowest terms. If θ is also an algebraic integer, then it is the root of a monic polynomial with rational integer coefficients, $\theta^n + a_1\theta^{n-1} + \cdots + a_n$, so we have an equation

$$(C/D)^n + a_1(C/D)^{n-1} + \cdots + a_{n-1}(C/D) + a_n = 0.$$

Multiplying through by D^n we get

$$C^n + a_1C^{n-1}D + \cdots + a_{n-1}CD^{n-1} + a_nD^n = 0,$$

where all terms are (ordinary) integers, and all but the first one is divisible by D . If $D > 1$ then it has some prime factor p , so all terms apart from the first are also divisible by p . Since the terms add up to zero, it follows that p divides C^n , which implies that p divides C , which contradicts the assertion that the fraction C/D is in its lowest terms. This in turn contradicts the hypothesis that θ can be expressed as a ratio of whole numbers in the first place. As the reader may like to verify, this fact implies the result attributed to Theaetetus above, that \sqrt{A} is irrational if and only if A is not a perfect square.

The *degree* of an algebraic number θ is defined to be the smallest degree, n , of any polynomial relation $\theta^n + a_1\theta^{n-1} + \cdots + a_{n-1}\theta + a_n = 0$ that θ satisfies, where the coefficients a_i are rational numbers. The corresponding polynomial, $P(X) = X^n + a_1X^{n-1} + \cdots + a_{n-1}X + a_n$ is unique, since if there were two of them then their difference would be of smaller degree and would also have θ as a root. (One could make it monic by dividing it through by the leading coefficient.) Let us call $P(X)$ the *minimal polynomial* of θ . The minimal polynomial is *irreducible* over the field of rational numbers: that is, it cannot be factored as a product of two polynomials, each of smaller degree and having rational numbers as coefficients. (If it could, then it would not be of minimal degree, since one of its factors would have θ as a root.) The minimal polynomial $P(X)$ of θ is a factor of any monic polynomial $G(X)$ with rational coefficients that has θ as root. (The greatest common divisor of P and G is another monic polynomial with rational coefficients that has θ as a root, so it cannot be of degree smaller than that of P and it must therefore be P .) The minimal polynomial $P(X)$ of θ has distinct roots. (If $P(X)$ had multiple roots, then a little elementary calculus shows that it would share a nontrivial factor with its derivative, $P'(X)$. Since the derivative is of lower degree than $P(X)$ and again has rational coefficients, the greatest common divisor of P and P' would provide a nontrivial factorization of $P(X)$, contradicting its irreducibility.)

A fundamental result due to Gauss is that the n th root of unity $\zeta_n = e^{2\pi i/n}$ is an algebraic integer of degree precisely $\phi(n)$, where ϕ is Euler's ϕ -function. For example, if p is prime, the minimal polynomial of ζ_p is

$$\frac{X^p - 1}{X - 1} = X^{p-1} + X^{p-2} + \cdots + X + 1,$$

which is of degree $\phi(p) = p - 1$.

15 Algebraic Numbers as Ciphers Determined by Their Minimal Polynomials

We have expressly insisted that our algebraic numbers are complex numbers (of a certain sort). But another possible attitude toward an algebraic number, θ , an attitude at times promoted by Kronecker, among others, is to deal with θ as an unknown satisfying only the algebraic relations implied by the fact that it is a root of its (unique monic) minimal polynomial with rational coefficients. For example, if the minimal polynomial of θ is $P(X) = X^3 - X - 1$, then, according to this view,

θ is just an algebraic symbol that comes with the rule that any occurrence of θ^3 may be replaced by $\theta + 1$ (rather as the complex number i can be regarded as a symbol with the property that i^2 may be replaced by -1). Any root of the minimal polynomial of θ satisfies all the same polynomial relations with rational coefficients that θ satisfies; these roots are called *conjugates* of θ . If θ is an algebraic number of degree n , then θ has n distinct conjugates, all of them again, of course, algebraic numbers.

16 A Few Remarks about the Theory of Polynomials

Central to the theory of polynomials in one variable—and, therefore, particularly to the theory of algebraic numbers—is the general relationship that *roots* have to *coefficients*:

$$\prod_{i=1}^n (X - T_i) = X^n + \sum_{j=0}^{n-1} (-1)^j A_j(T_1, T_2, \dots, T_n) X^{n-j}.$$

The polynomial $A_j(T_1, T_2, \dots, T_n)$ is homogeneous of degree j (this means that every monomial in it has total degree j), has integer coefficients, and is symmetric in (i.e., unchanged by any permutation of) the variables T_1, T_2, \dots, T_n .

The constant term is the product of the roots:

$$A_n(T_1, T_2, \dots, T_n) = T_1 \cdot T_2 \cdot \dots \cdot T_n,$$

which is known as the *norm* form. The coefficient of X^{n-1} is the sum of the roots:

$$A_1(T_1, T_2, \dots, T_n) = T_1 + T_2 + \dots + T_n,$$

and this is the *trace* form.

When $n = 2$ the norm and trace are all the symmetric polynomials in the list. For $n = 3$, beyond the norm and trace we also have the symmetric polynomial of degree two:

$$\begin{aligned} A_2(T_1, T_2, T_3) &= T_1 T_2 + T_2 T_3 + T_3 T_1 \\ &= \frac{1}{2} \{ (T_1 + T_2 + T_3)^2 - (T_1^2 + T_2^2 + T_3^2) \}. \end{aligned}$$

It is of major importance to this theory, and more specifically to GALOIS THEORY [V.24], that the symmetry properties of the conjugate roots are nicely reflected in these symmetric polynomials. In particular, we have the fundamental result that *any* symmetric polynomial in T_1, T_2, \dots, T_n with rational coefficients can be expressed as a polynomial with rational coefficients in the symmetric polynomials $A_j(T_1, T_2, \dots, T_n)$, and similarly with integral coefficients. For example, the equation above shows that $T_1^2 + T_2^2 + T_3^2$ can be expressed

as

$$A_1(T_1, T_2, T_3)^2 - 2A_2(T_1, T_2, T_3).$$

17 Fields of Algebraic Numbers and Rings of Algebraic Integers

The inverse of a nonzero algebraic number is again an algebraic number; the sum, difference, and product of two algebraic numbers are algebraic numbers; the sum, difference, and product of two algebraic integers are algebraic integers. The neat proofs of these (latter) facts are a good demonstration of the power of linear algebra, and in particular of *Cramer's rule*. This states that any matrix with integer coefficients (and therefore also any linear transformation of a finite-dimensional vector space that preserves an integer lattice) satisfies a monic polynomial identity with integer coefficients.

To see just how useful this remark is for finding polynomial relations, and more specifically for showing that the collections of algebraic numbers and algebraic integers are closed under sums and products, try your hand at showing that $\sqrt{2} + \sqrt{3}$ is an algebraic integer. One way to do it is to search for the monic fourth-degree polynomial equation that it satisfies. But this is hardly a beautiful calculation! If, however, you are familiar with linear algebra, then a less painful route is to form the four-dimensional vector space over the rational numbers, generated by $1, \sqrt{2}, \sqrt{3}$, and $\sqrt{6}$ (which are linearly independent when the scalars are rational). Multiplication by $\sqrt{2} + \sqrt{3}$ defines a linear transformation T of this vector space, and one can compute its characteristic polynomial P . The *Cayley-Hamilton theorem* says that $P(T) = 0$, and this translates into the statement that $\sqrt{2} + \sqrt{3}$ is a root of P .

These “closure properties” we have just discussed lead us to study, in complete generality, fields of algebraic numbers and rings of algebraic integers. A *number field* is a field that is generated (as a field) by finitely many algebraic numbers. A standard result tells us that any number field K can in fact be generated by a single carefully chosen algebraic number. The degree of this algebraic number equals the *degree* of K , which is defined to be the dimension of K when K is viewed as a vector space over the field \mathbb{Q} of rational numbers. One of the main introductory observations of Galois theory is that if K is a number field of degree n , then there are exactly n distinct ring homomorphisms (“embeddings”) $\iota : K \rightarrow \mathbb{C}$ from K into the field of complex numbers. (This means that ι sends 1 to 1 and respects

the addition and multiplication laws within K . That is, $\iota(x + y) = \iota(x) + \iota(y)$ and $\iota(x \cdot y) = \iota(x) \cdot \iota(y)$. From these imbeddings, we can construct some very useful rational-valued functions on K . For any element x in K , we form the n complex numbers x_1, x_2, \dots, x_n that are the images of x under the n different imbeddings of K into \mathbb{C} . We then let

$$a_j(x) = A_j(x_1, x_2, \dots, x_n),$$

where $A_j(X_1, X_2, \dots, X_n)$ is the j th symmetric polynomial of section 14 above. (Because the polynomials A_j are symmetric, we do not have to worry about the order of the images x_1, x_2, \dots, x_n in the above expression.) It is not immediately obvious that the values of a_j are rational numbers, but there is a theorem that tells us this.

If an algebraic number θ in K generates K (as a field), then the rational numbers $a_j(\theta)$ are the coefficients of its minimal polynomial; in general they are the coefficients of a power of its minimal polynomial. The most prominent of these functions are the multiplicative function $a_n(x) = x_1 \cdot x_2 \cdot \dots \cdot x_n$, called the *norm* function, usually denoted $x \mapsto N_{K/\mathbb{Q}}(x)$, and the additive function $a_1(x) = x_1 + x_2 + \dots + x_n$, called the *trace* function, usually denoted $x \mapsto \text{trace}_{K/\mathbb{Q}}(x)$.

The trace function can be used to define a fundamental symmetric bilinear form on the \mathbb{Q} -vector space K ,

$$\langle x, y \rangle = \text{trace}_{K/\mathbb{Q}}(x \cdot y),$$

which turns out to be nondegenerate. This nondegeneracy, together with the fact that if x, y are both algebraic integers, then $\langle x, y \rangle$ is an ordinary integer, can be used to show that the ring $\mathcal{O}(K)$ of *all* algebraic integers in K is finitely generated as an additive group. More specifically, there is a *basis* of algebraic integers in K , that is, a finite set $\{\theta_1, \theta_2, \dots, \theta_n\}$, such that any other algebraic integer in K can be expressed as an “ordinary” integer combination of the numbers θ_i .

Let us summarize this structure. The number field K is a finite-dimensional vector space over \mathbb{Q} and comes equipped with a nondegenerate bilinear symmetric form $(x, y) \mapsto \langle x, y \rangle$, and also with a lattice $\mathcal{O}(K) \subset K$. Moreover, the restriction of the bilinear form to $\mathcal{O}(K)$ takes on integral values.

The *discriminant* of K , denoted $D(K)$, is defined to be the DETERMINANT [III.15] of the matrix whose ij -entry is $\langle \theta_i, \theta_j \rangle$, for $\{\theta_1, \theta_2, \dots, \theta_n\}$ a basis of the lattice $\mathcal{O}(K)$; this determinant does not depend on the basis chosen.

The discriminant represents important information about the number field K . For one thing, there is a natural generalization to any number field of the notions of *splitting* and *ramification* that we discussed for quadratic fields, and the prime divisors p of $D(K)$ are precisely those prime numbers that ramify in the field extension K . By a theorem of MINKOWSKI [VI.64], the absolute value of the discriminant $D(K)$ of a number field K of degree n is always greater than

$$\left(\frac{\pi}{4}\right)^n \cdot \left(\frac{n^n}{n!}\right)^2.$$

This is greater than 1 unless K is the field of rational numbers. It follows that any nontrivial extension of the field of rational numbers has some prime that ramifies in it, a result that would be very hard to prove without the help of the algebraic structures we have just defined. This integer $D(K)$ really is quite a discriminating “tag” for our number field K , for, by a theorem of HERMITE [VI.47], given any integer D there are only finitely many different number fields with discriminant equal to D . (Not all integers can be discriminants: as is true for quadratic number fields, the integers D that are discriminants are either divisible by 4 or else congruent to 1 modulo 4.)

18 On the Size(s) of the Absolute Values of All Conjugates of an Algebraic Integer

As we have just seen, the coefficients of the minimal polynomial for an algebraic integer θ are given by the ordinary integers $a_j(\theta_1, \theta_2, \dots, \theta_n)$, where the numbers θ_i are all the conjugates of θ . The sizes of all these coefficients must therefore all be less than some universal number M that depends only on the degree of θ and the largest absolute value of any of its conjugates. As a consequence, given any n and any positive number B , there are only finitely many algebraic integers θ of degree less than n such that the absolute values of θ and its conjugates are all less than B . (This is because for any n and M there are only finitely many polynomials of degree less than or equal to n with the absolute values of all their integer coefficients at most M .) This finiteness result is the key to the following observation, due to Kronecker: if θ is an algebraic number and if the absolute values of θ and of all of its conjugates are equal to 1, then θ is a root of unity. Indeed, all the powers of θ have degree at most that of θ , and they enjoy the same property: their absolute value, and that of all their conjugates, is equal to 1. Consequently, there are only finitely many such algebraic numbers, from which

PUP: I can confirm that the use of ‘ $a_n(x)$ ’ and ‘ $a_1(x)$ ’ is OK.

it follows that there must be at least one coincidence of the form $\Theta^a = \Theta^b$ for different a and b . But this can happen only if Θ is a root of unity.

19 Weil Numbers

To follow this thread for just a bit, let us generalize the hypothesis of Kronecker's observation, and define a *Weil number*³ of absolute value r to be a nonzero algebraic integer such that it and all of its conjugates have the same absolute value r . By the discussion in the previous section there are only finitely many distinct Weil numbers of given degree and absolute value. By Kronecker's theorem, which we have just described, the Weil numbers of absolute value 1 are precisely the roots of unity. Here are further basic facts that you might try to prove. First, the quadratic Weil numbers ω are precisely those quadratic algebraic integers such that $|\text{trace}(\omega)| \leq 2\sqrt{|N(\omega)|} = 2\sqrt{|\omega\omega'|}$, where ω' is the (algebraic) conjugate of ω . Second, if p is prime then a quadratic Weil number ω of absolute value \sqrt{p} is a prime element of the (unique) ring of quadratic integers R_d that contains ω , and therefore gives a prime factorization $\omega\omega' = \pm p$ of the integer p in that ring.

Weil numbers of absolute value $p^{v/2}$, where p is again a prime number and v is a natural number, are extremely important in arithmetic: they hold the key to counting numbers of rational solutions of systems of polynomial equations over finite fields. For just one concrete example, the Gaussian integer $\omega = -1 + i$ and its algebraic conjugate (which, in this instance, is also its complex conjugate) $\bar{\omega} = -1 - i$ are Weil numbers (of absolute value 2) that control the number of solutions of the equation $y^2 - y = x^3 - x$ over all finite fields of size a power of 2. Specifically, the number of solutions of that equation over a field of order 2^v is given by the formula

$$2^v - (-1 - i)^v - (-1 + i)^v$$

(which is an ordinary integer). This leads to another immense chapter of mathematics.

20 Epilogue

The single symmetry $\alpha \mapsto \alpha'$, the algebraic conjugation in the rings R_d that we have discussed, gave birth, thanks to ABEL [VI.33] and GALOIS [VI.41] in the beginning of the nineteenth century, to the rich study of

(Galois) groups of symmetries of general number fields (see THE INSOLUBILITY OF THE QUINTIC [V.24]). This study continues with great intensity, since these Galois groups and their linear representations hold the key to a very detailed understanding of number fields. In its modern dress, algebraic number theory is closely connected with what is often called ARITHMETIC GEOMETRY [IV.5]. Kronecker's dream of getting explicit control of a wealth of algebraic number theoretic material by expressing algebraic numbers in terms of natural analytic functions has not yet been fully realized. Nevertheless, the scope of this dream (and, one might also add, the supply of natural analytic and algebraic functions) has expanded substantially: the full range of algebraic geometry and group representation theory is now being brought to bear on it. This is done, for example, by the *Langlands program*, which among other things works with objects known as *Shimura varieties*. On the one hand, these varieties have close connections with the theory of group representations and classical algebraic geometry, which greatly helps us to understand them. On the other hand, they are a rich source of concrete linear representations of Galois groups of number fields. This program, one of the glories of current mathematics, will, I expect, make a terrific chapter for a *Companion to Mathematics* to be written at the beginning of the next century.

Further reading

Basic Texts

First, I list three classics that require a minimum of background.

- Gauss, C. F. 1986. *Disquisitiones Arithmeticae*, English edn. New York: Springer.
- Davenport, H. 1992. *The Higher Arithmetic: An Introduction to the Theory of Numbers*. Cambridge: Cambridge University Press.
- Hardy, G. H., and E. M. Wright. 1979. *Introduction to Number Theory*. Oxford: Oxford University Press.

At a more advanced level, the following are extraordinary expository books.

- Borevich, Z. I., and I. R. Shafarevich. 1966. *Number Theory*. New York: Academic Press.
- Cassels, J., and A. Fröhlich. 1967. *Algebraic Number Theory*. New York: Academic Press.
- Cohen, H. 1993. *A Course in Computational Algebraic Number Theory*. New York: Springer.
- Ireland, K., and M. Rosen. 1982. *A Classical Introduction to Modern Number Theory*, 2nd edn. New York: Springer.
- Serre, J.-P. 1973. *A Course in Arithmetic*. New York: Springer.

3. This is a weaker condition than is usually required for Weil numbers but our deviation from standard usage should not be the cause of too much confusion.

Technical Articles and Books

- Baker, A. 1971. Imaginary quadratic fields with class number 2. *Annals of Mathematics* (2) 94:139–52.
- Brauer, R. 1950. On the Zeta-function of algebraic number fields. I. *American Journal of Mathematics* 69:243–50.
- Brauer, R. 1950. On the Zeta-function of algebraic number fields. II. *American Journal of Mathematics* 72:739–46.
- Goldfeld, D. 1985. Gauss's class number problem for imaginary quadratic fields. *Bulletin of the American Mathematical Society* 13:23–37.
- Gross, B., and D. Zagier. 1986. Heegner points and derivatives of L -series. *Inventiones Mathematicae* 84:225–320.
- Heegner, K. 1952. Diophantische Analysis und Modulfunktionen. *Mathematische Zeitschrift* 56:227–53.
- Hua, L.-K. 1942. On the least solution of Pell's equation. *Bulletin of the American Mathematical Society* 48:731–35.
- Lang, S. 1970. *Algebraic Number Theory*. Reading, MA: Addison-Wesley.
- Narkiewicz, W. 1973. *Algebraic Numbers*. Warsaw: Polish Scientific Publishers.
- Siegel, C. L. 1935. Über die Classenzahl quadratischer Zahlkörper. *Acta Arithmetica* 1:83–86.
- Stark, H. 1967. A complete determination of the complex quadratic fields of class-number one. *Michigan Mathematical Journal* 14:1–27.

IV.2 Analytic Number Theory

Andrew Granville

1 Introduction

What is number theory? One might have thought that it was simply the study of numbers, but that is too broad a definition, since numbers are almost ubiquitous in mathematics. To see what distinguishes number theory from the rest of mathematics, let us look at the equation $x^2 + y^2 = 15\,925$, and consider whether it has any solutions. One answer is that it certainly does: indeed, the solution set forms a circle of radius $\sqrt{15\,925}$ in the plane. However, a number theorist is interested in *integer* solutions, and now it is much less obvious whether any such solutions exist.

A useful first step in considering the above question is to notice that 15 925 is a multiple of 25: in fact, it is 25×637 . Furthermore, the number 637 can be decomposed further: it is 49×13 . That is, $15\,925 = 5^2 \times 7^2 \times 13$. This information helps us a lot, because if we can find integers a and b such that $a^2 + b^2 = 13$, then we can multiply them by $5 \times 7 = 35$ and we will have a solution to the original equation. Now we notice that $a = 2$ and $b = 3$ works, since $2^2 + 3^2 = 13$. Multiplying these numbers by 35, we obtain the solution $70^2 + 105^2 = 15\,925$ to the original equation.

As this simple example shows, it is often useful to decompose positive integers multiplicatively into components that cannot be broken down any further. These components are called *prime numbers*, and THE FUNDAMENTAL THEOREM OF ARITHMETIC [V.16] states that every positive integer can be written as a product of primes in exactly one way. That is, there is a one-to-one correspondence between positive integers and finite products of primes. In many situations we know what we need to know about a positive integer once we have decomposed it into its prime factors and understood those, just as we can understand a lot about molecules by studying the atoms of which they are composed. For example, it is known that the equation $x^2 + y^2 = n$ has an integer solution if and only if every prime of the form $4m+3$ occurs an even number of times in the prime factorization of n . (This tells us, for instance, that there are no integer solutions to the equation $x^2 + y^2 = 13\,475$, since $13\,475 = 5^2 \times 7^2 \times 11$, and 11 appears an odd number of times in this product.)

Once one begins the process of determining which integers are primes and which are not, it is soon apparent that there are many primes. However, as one goes further and further, the primes seem to consist of a smaller and smaller proportion of the positive integers. They also seem to come in a somewhat irregular pattern, which raises the question of whether there is any formula that describes all of them. Failing that, can one perhaps describe a large class of them? We can also ask whether there are infinitely many primes. If there are, can we quickly determine how many there are up to a given point? Or at least give a good estimate for this number? Finally, when one has spent long enough looking for primes, one cannot help but ask whether there is a quick way of recognizing them. This last question is discussed in COMPUTATIONAL NUMBER THEORY [IV.3]; the rest motivate the present article.

Now that we have discussed what marks number theory out from the rest of mathematics, we are ready to make a further distinction: between *algebraic* and *analytic* number theory. The main difference is that in algebraic number theory (which is the main topic of ALGEBRAIC NUMBERS [IV.1]) one typically considers questions with answers that are given by exact formulas, whereas in analytic number theory, the topic of this article, one looks for *good approximations*. For the sort of quantity that one estimates in analytic number theory, one does not expect an exact formula to exist, except perhaps one of a rather artificial and unilluminating kind. One of the best examples of such a

quantity is one we shall discuss in detail: the number of primes less than or equal to x .

Since we are discussing approximations, we will need terminology that allows us to give some idea of the quality of an approximation. Suppose, for example, that we have a rather erratic function $f(x)$ but are able to show that, once x is large enough, $f(x)$ is never bigger than $25x^2$. This is useful because we understand the function $g(x) = x^2$ quite well. In general, if we can find a constant c such that $|f(x)| \leq cg(x)$ for every x , then we write $f(x) = O(g(x))$. A typical usage occurs in the sentence “the average number of prime factors of an integer up to x is $\log \log x + O(1)$ ”; in other words, there exists some constant $c > 0$ such that $|\text{the average} - \log \log x| \leq c$ once x is sufficiently large.

We write $f(x) \sim g(x)$ if $\lim_{x \rightarrow \infty} f(x)/g(x) = 1$; and also $f(x) \approx g(x)$ when we are being a little less precise, that is, when we want to say that $f(x)$ and $g(x)$ come close when x is sufficiently large, but we cannot be, or do not want to be, more specific about what we mean by “come close.”

It is convenient for us to use the notation \sum for sums and \prod for product. Typically we will indicate beneath the symbol what terms the sum, or product, is to be taken over. For example, $\sum_{m \geq 2}$ will be a sum over all integers m that are greater than or equal to 2, whereas $\prod_{p \text{ prime}}$ will be a product over all primes p .

2 Bounds for the Number of Primes

Ancient Greek mathematicians knew that there were infinitely many primes. Their beautiful proof by contradiction goes as follows. Suppose that there are only finitely many primes, say k of them, which we will denote by p_1, p_2, \dots, p_k . What are the prime factors of $p_1 p_2 \cdots p_k + 1$? Since this number is greater than 1 it must have at least one prime factor, and this must be p_j for some j (since *all* primes are contained among p_1, p_2, \dots, p_k). But then p_j divides both $p_1 p_2 \cdots p_k$ and $p_1 p_2 \cdots p_k + 1$, and hence their difference, 1, which is impossible.

Many people dislike this proof, since it does not actually exhibit infinitely many primes: it merely shows that there cannot be finitely many. It is more or less possible to correct this deficiency by defining the sequence $x_1 = 2$, $x_2 = 3$, and $x_{k+1} = x_1 x_2 \cdots x_k + 1$ for each $k \geq 2$. Then each x_k must contain at least one prime factor, q_k say, and these prime factors must be distinct, since if $k < \ell$, then q_k divides x_k which divides $x_\ell - 1$,

while q_ℓ divides x_ℓ . This gives us an infinite sequence of primes.

In the eighteenth century EULER [VI.19] gave a different proof that there are infinitely many primes, one that turned out to be highly influential in what was to come later. Suppose again that the list of primes is p_1, p_2, \dots, p_k . As we have mentioned, the fundamental theorem of arithmetic implies that there is a one-to-one correspondence between the set of all integers and the set of products of the primes, which, if those are the only primes, is the set $\{p_1^{a_1} p_2^{a_2} \cdots p_k^{a_k} : a_1, a_2, \dots, a_k \geq 0\}$. But, as Euler observed, this implies that a sum involving the elements of the first set should equal the analogous sum involving the elements of the second set:

$$\begin{aligned} & \sum_{\substack{n \geq 1 \\ n \text{ a positive integer}}} \frac{1}{n^s} \\ &= \sum_{a_1, a_2, \dots, a_k \geq 0} \frac{1}{(p_1^{a_1} p_2^{a_2} \cdots p_k^{a_k})^s} \\ &= \left(\sum_{a_1 \geq 0} \frac{1}{(p_1^{a_1})^s} \right) \left(\sum_{a_2 \geq 0} \frac{1}{(p_2^{a_2})^s} \right) \cdots \left(\sum_{a_k \geq 0} \frac{1}{(p_k^{a_k})^s} \right) \\ &= \prod_{j=1}^k \left(1 - \frac{1}{p_j^s} \right)^{-1}. \end{aligned}$$

The last equality holds because each sum in the second-last line is the sum of a geometric progression. Euler then noted that if we take $s = 1$, the right-hand side equals some rational number (since each $p_j > 1$) whereas the left-hand side equals ∞ . This is a contradiction, so there cannot be finitely many primes. (To see why the left-hand side is infinite when $s = 1$, note that $(1/n) \geq \int_n^{n+1} (1/t) dt$ since the function $1/t$ is decreasing, and therefore $\sum_{n=1}^{N-1} (1/n) \geq \int_1^N (1/t) dt = \log N$ which tends to ∞ as $N \rightarrow \infty$.)

During the proof above, we gave a formula for $\sum n^{-s}$ under the false assumption that there are only finitely many primes. To correct it, all we have to do is rewrite it in the obvious way without that assumption:

$$\sum_{\substack{n \geq 1 \\ n \text{ a positive integer}}} \frac{1}{n^s} = \prod_{p \text{ prime}} \left(1 - \frac{1}{p^s} \right)^{-1}. \quad (1)$$

Now, however, we need to be a little careful about whether the two sides of the formula converge. It is safe to write down such a formula when both sides are absolutely convergent, and this is true when $s > 1$. (An infinite sum or product is *absolutely convergent* if the value does not change when we take the terms in any order we want.)

PUP: point taken about 'beneath' not being entirely unproblematic, but this couldn't confuse anyone who has read this far and is true most of the time (when these symbols appear in displays), so we'd prefer to leave it as it is. OK?

Like Euler, we want to be able to interpret what happens to (1) when $s = 1$. Since both sides converge and are equal when $s > 1$, the natural thing to do is consider their common limit as s tends to 1 from above. To do this we note, as above, that the left-hand side of (1) is well approximated by

$$\int_1^\infty \frac{dt}{t^s} = \frac{1}{s-1},$$

so it diverges as $s \rightarrow 1^+$. We deduce that

$$\prod_{p \text{ prime}} \left(1 - \frac{1}{p}\right) = 0. \quad (2)$$

Upon taking logarithms and discarding negligible terms, this implies that

$$\sum_{p \text{ prime}} \frac{1}{p} = \infty. \quad (3)$$

So how numerous are the primes? One way to get an idea is to determine the behavior of the sum analogous to (3) for other sequences of integers. For instance, $\sum_{n \geq 1} 1/n^2$ converges, so the primes are, in this sense, more numerous than the squares. This argument works if we replace the power 2 by any $s > 1$, since then, as we have just observed, the sum $\sum_{n \geq 1} 1/n^s$ is about $1/(s-1)$ and in particular converges. In fact, since $\sum_{n \geq 1} 1/n(\log n)^2$ converges, we see that the primes are in the same sense more numerous than the numbers $\{n(\log n)^2 : n \geq 1\}$, and hence there are infinitely many integers x for which the number of primes less than or equal to x is at least $x/(\log x)^2$.

Thus, there seem to be primes in abundance, but we would also like to verify our observations, made from calculations, that the primes constitute a smaller and smaller proportion of the integers as the integers become larger and larger. The easiest way to see this is to try to count the primes using the “sieve of Eratosthenes.” In the sieve of Eratosthenes one starts with all the positive integers up to some number x . From these, one deletes the numbers 4, 6, 8 and so on—that is, all multiples of 2 apart from 2 itself. One then takes the first undeleted integer greater than 2, which is 3, and deletes all its multiples—again, not including the number 3 itself. Then one removes all multiples of 5 apart from 5, and so on. By the end of this process, one is left with the primes up to x .

This suggests a way to guess at how many there are. After deleting every second integer up to x other than 2 (which we call “sieving by 2”) one is left with roughly half the integers up to x ; after sieving by 3, one is left with roughly two thirds of those that had remained;

continuing like this we expect to have about

$$x \prod_{p \leq y} \left(1 - \frac{1}{p}\right) \quad (4)$$

integers left by the time we have sieved with all the primes up to y . Once $y = \sqrt{x}$ the undeleted integers are 1 and the primes up to x , since every composite has a prime factor no bigger than its square root. So, is (4) a good approximation for the number of primes up to x when $y = \sqrt{x}$?

To answer this question, we need to be more precise about what the formula in (4) is estimating. It is supposed to approximate the number of integers up to x that have no prime factors less than or equal to y , plus the number of primes up to y . The so-called *inclusion-exclusion principle* can be used to show that the approximation given in (4) is accurate to within 2^k , where k is the number of primes less than or equal to y . Unless k is very small, this error term of 2^k is far larger than the quantity we are trying to estimate, and the approximation is useless. It is quite good if k is less than a small constant times $\log x$, but, as we have seen, this is far less than the number of primes we expect up to y if $y \approx \sqrt{x}$. Thus it is not clear whether (4) can be used to obtain a good estimate for the number of primes up to x . What we *can* do, however, is use this argument to give an upper bound for the number of primes up to x , since the number of primes up to x is never more than the number of integers up to x that are free of prime factors less than or equal to y , plus the number of primes up to y , which is no more than 2^k plus the expression in (4).

Now, by (2), we know that as y gets larger and larger the product $\prod_{p \leq y} (1 - 1/p)$ converges to zero. Therefore, for any small positive number ε we can find a y such that $\prod_{p \leq y} (1 - 1/p) < \varepsilon/2$. Since every term in this product is at least $1/2$, the product is at least $1/2^k$. Hence, for any $x \geq 2^{2k}$ our error term, 2^k , is no bigger than the quantity in (4), and therefore the number of primes up to x is no larger than twice (4), which, by our choice of y , is less than εx . Since we were free to make ε as small as we liked, the primes are indeed a vanishing proportion of all the integers, as we predicted.

Even though the error term in the inclusion-exclusion principle is too large for us to use that method to estimate (4) when $y = \sqrt{x}$, we can still hope that (4) is a good approximation for the number of primes up to x : perhaps a different argument would give us a much smaller error term. And this turns out to be the case: in fact, the error never gets much bigger than (4). However, when $y = \sqrt{x}$ the number of primes up to x is

actually about 8/9 times (4). So why does (4) not give a good approximation? After sieving with prime p we supposed that roughly 1 in every p of the remaining integers were deleted: a careful analysis yields that this can be justified when p is small, but that this becomes an increasingly poor approximation of what really happens for larger p ; in fact (4) *does not* give a correct approximation once y is bigger than a fixed power of x . So what goes wrong? In the hope that the proportion is roughly $1/p$ lies the unspoken assumption that the consequences of sieving by p are independent of what happened with the primes smaller than p . But if the primes under consideration are no longer small, then this assumption is false. This is one of the main reasons that it is hard to estimate the number of primes up to x , and indeed similar difficulties lie at the heart of many related problems.

One can refine the bounds given above but they do not seem to yield an asymptotic estimate for the primes (that is, an estimate which is correct to within a factor that tends to 1 as x gets large). The first good guesses for such an estimate emerged at the beginning of the nineteenth century, none better than what emerges from an observation of GAUSS [VI.26], made when studying tables of primes up to three million at sixteen years of age, that “the density of primes at around x is about $1/\log x$.” Interpreting this, we guess that the number of primes up to x is about

$$\sum_{n=2}^x \frac{1}{\log n} \approx \int_2^x \frac{dt}{\log t}.$$

Let us compare this prediction (rounded to the nearest integer) with the latest data on numbers of primes, discovered by a mixture of ingenuity and computational power. Table 1 shows the actual numbers of primes up to various powers of 10 together with the difference between these numbers and what Gauss’s formula gives. The differences are far smaller than the numbers themselves, so his prediction is amazingly accurate. It does seem always to be an overcount, but since the width of the last column is about half that of the central one it appears that the difference is something like \sqrt{x} .

In the 1930s, the great probability theorist, Cramér, gave a probabilistic way of interpreting Gauss’s prediction. We can represent the primes as a sequence of 0s and 1s: Putting a “1” each time we encounter a prime, and a “0” otherwise, we obtain, starting from 3, the sequence 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, Cramér’s idea is to suppose that this sequence, which represents the

Table 1 Primes up to various x , and the overcount in Gauss’s prediction.

x	$\pi(x) = \#\{\text{primes} \leq x\}$	Overcount: $\int_2^x \frac{dt}{\log t} - \pi(x)$
10^8	5 761 455	753
10^9	50 847 534	1 700
10^{10}	455 052 511	3 103
10^{11}	4 118 054 813	11 587
10^{12}	37 607 912 018	38 262
10^{13}	346 065 536 839	108 970
10^{14}	3 204 941 750 802	314 889
10^{15}	29 844 570 422 669	1 052 618
10^{16}	279 238 341 033 925	3 214 631
10^{17}	2 623 557 157 654 233	7 956 588
10^{18}	24 739 954 287 740 860	21 949 554
10^{19}	234 057 667 276 344 607	99 877 774
10^{20}	2 220 819 602 560 918 840	222 744 643
10^{21}	21 127 269 486 018 731 928	597 394 253
10^{22}	201 467 286 689 315 906 290	1 932 355 207

primes, has the same properties as a “typical” sequence of 0s and 1s, and to use this principle to make precise conjectures about the primes. More precisely, let X_3, X_4, \dots be an infinite sequence of RANDOM VARIABLES [III.73 §4] taking the values 0 or 1, and let the variable X_n equal 1 with probability $1/\log n$ (so that it equals 0 with probability $1 - 1/\log n$). Assume also that the variables are independent, so for each m knowledge about the variables other than X_m tells us nothing about X_m itself. Cramér’s suggestion was that any statement about the distribution of 1s in the sequence that represents the primes will be true if and only if it is true with probability 1 for his random sequences. Some care is needed in interpreting this statement: for example, with probability 1 a random sequence will contain infinitely many even numbers. However, it is possible to formulate a general principle that takes account of such examples.

Here is an example of a use of the Gauss–Cramér model. With the help of the CENTRAL LIMIT THEOREM [III.73 §5] one can prove that, with probability 1, there are

$$\int_2^x \frac{dt}{\log t} + O(\sqrt{x} \log x)$$

1s among the first x terms in our sequence. The model tells us that the same should be true of the sequence representing primes, and so we predict that

$$\#\{\text{primes up to } x\} = \int_2^x \frac{dt}{\log t} + O(\sqrt{x} \log x), \quad (5)$$

just as the table suggests.

The Gauss–Cramér model provides a beautiful way to think about distribution questions concerning the prime numbers, but it does not give proofs, and it does not seem likely that it can be made into such a tool; so for proofs we must look elsewhere. In analytic number theory one attempts to count objects that appear naturally in arithmetic, yet which resist being counted easily. So far, our discussion of the primes has concentrated on upper and lower bounds that follow from their basic definition and a few elementary properties—notably the fundamental theorem of arithmetic. Some of these bounds are good and some not so good. To improve on these bounds we shall do something that seems unnatural at first, and reformulate our question as a question about complex functions. This will allow us to draw on deep tools from analysis.

3 The “Analysis” in Analytic Number Theory

These analytic techniques were born in an 1859 memoir of RIEMANN [VI.49], in which he looked at the function that appears in the formula (1) of Euler, but with one crucial difference: now he considered *complex* values of s . To be precise, he defined what we now call the *Riemann zeta function* as follows:

$$\zeta(s) = \sum_{n \geq 1} \frac{1}{n^s}.$$

It can be shown quite easily that this sum converges whenever the real part of s is greater than 1, as we have already seen in the case of real s . However, one of the great advantages of allowing complex values of s is that the resulting function is *HOLOMORPHIC* [I.3 §5.6], and we can use a process of *analytic continuation* to make sense of $\zeta(s)$ for every s apart from 1. (A similar but more elementary example of this phenomenon is the infinite series $\sum_{n \geq 0} z^n$, which converges if and only if $|z| < 1$. However, when it does converge, it equals $1/(1 - z)$, and this formula defines a holomorphic function that is defined everywhere except $z = 1$.) Riemann proved the remarkable fact that confirming Gauss’s conjecture for the number of primes up to x is equivalent to gaining a good understanding of the zeros of the function $\zeta(s)$, that is, of the values of s for which $\zeta(s) = 0$. Riemann’s deep work gave birth to our subject, so it seems worthwhile to at least sketch the key steps in the argument linking these seemingly unconnected topics.

Riemann’s starting point was Euler’s formula (1). It is not hard to prove that this formula is valid when s is

complex, as long as its real part is greater than 1, so we have

$$\zeta(s) = \prod_{p \text{ prime}} \left(1 - \frac{1}{p^s}\right)^{-1}.$$

If we take the logarithm of both sides and then differentiate, we obtain the equation

$$-\frac{\zeta'(s)}{\zeta(s)} = \sum_{p \text{ prime}} \frac{\log p}{p^s - 1} = \sum_{p \text{ prime}} \sum_{m \geq 1} \frac{\log p}{p^{ms}}.$$

We need some way to distinguish between primes $p \leq x$ and primes $p > x$; that is, we want to count those primes p for which $x/p \geq 1$, but not those with $x/p < 1$. This can be done using the *step function* that takes the value 0 for $y < 1$ and the value 1 for $y > 1$ (so that its graph looks like a step). At $y = 1$, the point of discontinuity, it is convenient to give the function the average value, $\frac{1}{2}$. Perron’s formula, one of the big tools of analytic number theory, describes this step function by an integral, as follows. For any $c > 0$,

$$\frac{1}{2\pi i} \int_{s: \operatorname{Re}(s)=c} \frac{y^s}{s} ds = \begin{cases} 0 & \text{if } 0 < y < 1, \\ \frac{1}{2} & \text{if } y = 1, \\ 1 & \text{if } y > 1. \end{cases}$$

The integral is a *path integral* along a vertical line in the complex plane: the line consisting of all points $c + it$ with $t \in \mathbb{R}$. We apply Perron’s formula with $y = x/p^m$, so that we count the term corresponding to p^m when $p^m < x$, but not when $p^m > x$. To avoid the “ $\frac{1}{2}$,” assume that x is not a prime power. In that case we obtain

$$\begin{aligned} & \sum_{\substack{p \text{ prime}, m \geq 1 \\ p^m \leq x}} \log p \\ &= \frac{1}{2\pi i} \sum_{p \text{ prime}, m \geq 1} \log p \int_{s: \operatorname{Re}(s)=c} \left(\frac{x}{p^m}\right)^s \frac{ds}{s} \\ &= -\frac{1}{2\pi i} \int_{s: \operatorname{Re}(s)=c} \frac{\zeta'(s)}{\zeta(s)} \frac{x^s}{s} ds. \end{aligned} \quad (6)$$

We can justify swapping the order of the sum and the integral if c is taken large enough, since everything then converges absolutely. Now the left-hand side of the above equation is not counting the number of primes up to x but rather a “weighted” version: for each prime p we add a weight of $\log p$ to the count. It turns out, though, that Gauss’s prediction for the number of primes up to x follows so long as we can show that x is a good estimate for this weighted count when x is large. Notice that the sum in (6) is exactly the logarithm of the lowest common multiple of the integers less than or equal to x , which perhaps explains why

this weighted counting function for the primes is a natural function to consider. Another explanation is that if the density of primes near p is indeed about $1/\log p$, then multiplying by a weight of $\log p$ makes the density everywhere about 1.

If you know some complex analysis, then you will know that *Cauchy's residue theorem* allows one to evaluate the integral in (6) in terms of the “residues” of the integrand $(\zeta'(s)/\zeta(s))(x^s/s)$, that is, the poles of this function. Moreover, for any function f that is analytic except perhaps at finitely many points, the poles of $f'(s)/f(s)$ are the zeros and poles of f . Each pole of $f'(s)/f(s)$ has order 1, and the residue is simply the order of the corresponding zero, or minus the order of the corresponding pole, of f . Using these facts we can obtain the *explicit formula*

$$\sum_{\substack{p \text{ prime}, m \geq 1 \\ p^m \leq x}} \log p = x - \sum_{\rho: \zeta(\rho)=0} \frac{x^\rho}{\rho} - \frac{\zeta'(0)}{\zeta(0)}. \quad (7)$$

Here the zeros of $\zeta(s)$ are counted with multiplicity: that is, if ρ is a zero of $\zeta(s)$ of order k , then there are k terms for ρ in the sum. It is astonishing that there can be such a formula, an exact expression for the number of primes up to x in terms of the zeros of a complicated function: you can see why Riemann's work stretched people's imagination and had such an impact.

Riemann made another surprising observation which allows us to easily determine the values of $\zeta(s)$ on the left-hand side of the complex plane (where the function is not naturally defined). The idea is to multiply $\zeta(s)$ by some simple function so that the resulting product $\xi(s)$ satisfies the *functional equation*

$$\xi(s) = \xi(1-s) \quad \text{for all } s. \quad (8)$$

He determined that this can be done by taking $\xi(s) = \frac{1}{2}s(s-1)\pi^{-s/2}\Gamma(\frac{1}{2}s)\zeta(s)$. Here $\Gamma(s)$ is the famous **GAMMA FUNCTION** [III.31], which equals the factorial function at positive integers (that is, $\Gamma(n) = (n-1)!$), and is well-defined and continuous for all other s .

A careful analysis of (1) reveals that there are no zeros of $\zeta(s)$ with $\text{Re}(s) > 1$. Then, with the help of (8), we can deduce that the only zeros of $\zeta(s)$ with $\text{Re}(s) < 0$ lie at the negative even integers $-2, -4, \dots$ (the “trivial zeros”). So, to be able to use (7), we need to determine the zeros inside the *critical strip*, the set of all s such that $0 \leq \text{Re}(s) \leq 1$. Here Riemann made yet another extraordinary observation which, if true, would allow us tremendous insight into virtually every aspect of the distribution of primes.

The Riemann hypothesis. If $0 \leq \text{Re}(s) \leq 1$ and $\zeta(s) = 0$, then $\text{Re}(s) = \frac{1}{2}$.

It is known that there are infinitely many zeros on the line $\text{Re}(s) = \frac{1}{2}$, crowding closer and closer together as we go up the line. The Riemann hypothesis has been verified computationally for the ten billion zeros of lowest height (that is, with $|\text{Im}(s)|$ smallest), it can be shown to hold for at least 40% of all zeros, and it fits nicely with many different heuristic assertions about the distribution of primes and other sequences. Yet, for all that, it remains an unproved hypothesis, perhaps the most famous and tantalizing in all of mathematics.

How did Riemann think of his “hypothesis”? Riemann's memoir gives no hint as to how he came up with such an extraordinary conjecture, and for a long time afterwards it was held up as an example of the great heights to which humankind could ascend by pure thought alone. However, in the 1920s Siegel and WEIL [VI.93] got hold of Riemann's unpublished notes and from these it is evident that Riemann had been able to determine the lowest few zeros to several decimal places through extensive hand calculations—so much for “pure thought alone”! Nevertheless, the Riemann hypothesis is a mammoth leap of imagination and to have come up with an algorithm to calculate zeros of $\zeta(s)$ is a remarkable achievement. (See **COMPUTATIONAL NUMBER THEORY** [IV.3] for a discussion of how zeros of $\zeta(s)$ can be calculated.)

If the Riemann hypothesis is true, then it is not hard to prove the bound

$$\left| \frac{x^\rho}{\rho} \right| \leq \frac{x^{1/2}}{|\text{Im}(\rho)|}.$$

Inserting this into (7) one can deduce that

$$\sum_{\substack{p \text{ prime} \\ p \leq x}} \log p = x + O(\sqrt{x} \log^2 x). \quad (9)$$

This, in turn, can be “translated” into (5). In fact these estimates hold if and only if the Riemann hypothesis is true.

The Riemann hypothesis is not an easy thing to understand, nor to fully appreciate. The equivalent, (5), is perhaps easier. Another version, which I prefer, is that, for every $N \geq 100$,

$$|\log(\text{lcm}[1, 2, \dots, N]) - N| \leq \sqrt{N}(\log N)^2.$$

To focus on the overcount in Gauss's guesstimate for the number of primes up to x , we use the following approximation, which can be deduced from (7) if, and

only if, the Riemann hypothesis is true:

$$\frac{\int_2^x (1/\log t) dt - \#\{\text{primes} \leq x\}}{\sqrt{x}/\log x} \approx 1 + 2 \sum_{\substack{\text{all real numbers } \gamma > 0 \\ \text{such that } \frac{1}{2} + i\gamma \\ \text{is a zero of } \zeta(s)}} \frac{\sin(\gamma \log x)}{\gamma}. \quad (10)$$

The right-hand side here is the overcount in Gauss's prediction for the number of primes up to x , divided by something that grows like \sqrt{x} . When we looked at the table of primes it seemed that this quantity should be roughly constant. However, that is not quite true as we see upon examining the right-hand side. The first term on the right-hand side, the "1," corresponds to the contribution of the squares of the primes in (7). The subsequent terms correspond to the terms involving the zeros of $\zeta(s)$ in (7); these terms have denominator γ so the most significant terms in this sum are those with the smallest values of γ . Moreover, each of these terms is a sine wave, which oscillates, half the time positive and half the time negative. Having the "log x " in there means that these oscillations happen slowly (which is why we hardly notice them in the table above), but they do happen, and indeed the quantity in (10) does eventually get negative. No one has yet determined a value of x for which this is negative (that is, a value of x for which there are more than $\int_2^x (1/\log t) dt$ primes up to x), though our best guess is that the first time this happens is for

$$x \approx 1.398 \times 10^{316}.$$

How does one arrive at such a guess given that the table of primes extends only up to 10^{22} ? One begins by using the first thousand terms of the right-hand side of (10) to approximate the left-hand side; wherever it looks as though it could be negative, one approximates with more terms, maybe a million, until one becomes pretty certain that the value is indeed negative.

It is not uncommon to try to understand a given function better by representing it as a sum of sines and cosines like this; indeed this is how one studies the harmonics in music, and (10) becomes quite compelling from this perspective. Some experts suggest that (10) tells us that "the primes have music in them" and thus makes the Riemann hypothesis believable, even desirable.

To prove unconditionally that

$$\#\{\text{primes} \leq x\} \sim \int_2^x \frac{dt}{\log t},$$

the so-called *prime number theorem*, we can take the same approach as above but, since we are not asking for such a strong approximation to the number of primes up to x , we need to show only that the zeros near to the line $\text{Re}(s) = 1$ do not contribute much to the formula (7). By the end of the nineteenth century this task had been reduced to showing that there are no zeros actually *on* the line $\text{Re}(s) = 1$: this was eventually established by DE LA VALLÉE POUSSIN [VI.67] and HADAMARD [VI.65] in 1896.

Subsequent research has provided wider and wider subregions of the critical strip without zeros of $\zeta(s)$ (and thus improved approximations to the number of primes up to x), without coming anywhere near to proving the Riemann hypothesis. This remains as an outstanding open problem of mathematics.

A simple question like "How many primes are there up to x ?" deserves a simple answer, one that uses elementary methods rather than all of these methods of complex analysis, which seem far from the question at hand. However, (7) tells us that the prime number theorem is true *if and only if* there are no zeros of $\zeta(s)$ on the line $\text{Re}(s) = 1$, and so one might argue that it is inevitable that complex analysis must be involved in such a proof. In 1949 Selberg and Erdős surprised the mathematical world by giving an elementary proof of the prime number theorem. Here, the word "elementary" does not mean "easy" but merely that the proof does not use advanced tools such as complex analysis—in fact, their argument is a complicated one. Of course their proof must somehow show that there is no zero on the line $\text{Re}(s) = 1$, and indeed their combinatorics cunningly masks a subtle complex analysis proof beneath the surface (read Ingham's discussion (1949) for a careful examination of the argument).

4 Primes in Arithmetic Progressions

After giving good estimates for the number of primes up to x , which from now on we shall denote by $\pi(x)$, we might ask for the number of such primes that are congruent to $a \bmod q$. (If you do not know what this means, see MODULAR ARITHMETIC [III.60].) Let us write $\pi(x; q, a)$ for this quantity. To start with, note that there is only one prime congruent to $2 \bmod 4$, and indeed there can be no more than one prime in any arithmetic $a, a + q, a + 2q, \dots$ if a and q have a common factor greater than 1. Let $\phi(q)$ denote the number of integers a , $1 \leq a \leq q$, such that $(a, q) = 1$. (The notation (a, q) stands for the highest common factor

of a and q .) Then all but a small finite number of the infinitely many primes belong to the $\phi(q)$ arithmetic progressions $a, a+q, a+2q, \dots$ with $1 \leq a < q$ and $(a, q) = 1$. Calculation reveals that the primes seem to be pretty evenly split between these $\phi(q)$ arithmetic progressions, so we might guess that in the limit the proportion of primes in each of them is $1/\phi(q)$. That is, whenever $(a, q) = 1$, we might conjecture that, as $x \rightarrow \infty$,

$$\pi(x; q, a) \sim \frac{\pi(x)}{\phi(q)}. \quad (11)$$

It is far from obvious even that the number of primes congruent to $a \pmod q$ is infinite. This is a famous theorem of DIRICHLET [VI.36]. To begin to consider such questions we need a systematic way to identify integers n that are congruent to $a \pmod q$, and this Dirichlet provided by introducing a class of functions now known as (*Dirichlet*) *characters*. Formally, a *character* mod q is a function χ from \mathbb{Z} to \mathbb{C} with the following three properties (in ascending order of interest):

- (i) $\chi(n) = 0$ whenever n and q have a common factor greater than 1;
- (ii) χ is *periodic* mod q (that is, $\chi(n+q) = \chi(n)$ for every integer n);
- (iii) χ is *multiplicative* (that is, $\chi(mn) = \chi(m)\chi(n)$ for any two integers m and n).

An easy but important example of a character mod q is the *principal character* χ_q , which takes the value 1 if $(n, q) = 1$ and 0 otherwise. If q is prime, then another important example is the *Legendre symbol* $(\frac{\cdot}{q})$: one sets $(\frac{n}{q})$ to be 0 if n is a multiple of q , 1 if n is a quadratic residue mod q , and -1 if n is a quadratic nonresidue mod q . (An integer n is called a *quadratic residue* mod q if n is congruent mod q to a perfect square.) If q is composite, then a function known as the *Legendre-Jacobi symbol* $(\frac{\cdot}{q})$, which generalizes the Legendre symbol, is also a character. This too is an important example that helps us, in a slightly less direct way, to recognize squares mod q .

These characters are all real-valued, which is the exception rather than the rule. Here is an example of a genuinely complex-valued character in the case $q = 5$. Set $\chi(n)$ to be 0 if $n \equiv 0 \pmod 5$, i if $n \equiv 2, -1$ if $n \equiv 4$, $-i$ if $n \equiv 3$, and 1 if $n \equiv 1$. To see that this is a character, note that the powers of 2 mod 5 are 2, 4, 3, 1, 2, 4, 3, 1, \dots , while the powers of i are $i, -1, -i, 1, i, -1, -i, 1, \dots$

It can be shown that there are precisely $\phi(q)$ distinct characters mod q . Their usefulness to us comes from

the properties above, together with the following formula, in which the sum is over all characters mod q and $\bar{\chi}(a)$ denotes the complex conjugate of $\chi(a)$:

$$\frac{1}{\phi(q)} \sum_{\chi} \bar{\chi}(a) \chi(n) = \begin{cases} 1 & \text{if } n \equiv a \pmod q, \\ 0 & \text{otherwise.} \end{cases}$$

What is this formula doing for us? Well, understanding the set of integers congruent to $a \pmod q$ is equivalent to understanding the function that takes the value 1 if $n \equiv a \pmod q$ and 0 otherwise. This function appears on the right-hand side of the formula. However, it is not a particularly nice function to deal with, so we write it as a linear combination of characters, which are much nicer functions because they are multiplicative. The coefficient associated with the character χ in this linear combination is the number $\bar{\chi}(a)/\phi(q)$.

From the formula, it follows that

$$\begin{aligned} & \sum_{\substack{p \text{ prime, } m \geq 1 \\ p^m \leq x \\ p^m \equiv a \pmod q}} \log p \\ &= \frac{1}{\phi(q)} \sum_{\chi \pmod q} \bar{\chi}(a) \sum_{\substack{p \text{ prime, } m \geq 1 \\ p^m \leq x}} \chi(p^m) \log p. \end{aligned}$$

The sum on the left-hand side is a natural adaptation of the sum we considered earlier when we were counting all primes. And we can estimate it if we can get good estimates for each of the sums

$$\sum_{\substack{p \text{ prime, } m \geq 1 \\ p^m \leq x}} \chi(p^m) \log p.$$

We approach these sums much as we did before, obtaining an explicit formula, analogous to (7), (10), now in terms of the zeros of the *Dirichlet L-function*:

$$L(s, \chi) = \sum_{n \geq 1} \frac{\chi(n)}{n^s}.$$

This function turns out to have properties closely analogous to the main properties of $\zeta(s)$. In particular, it is here that the multiplicativity of χ is all-important, since it gives us a formula similar to (1):

$$\sum_{n \geq 1} \frac{\chi(n)}{n^s} = \prod_{p \text{ prime}} \left(1 - \frac{\chi(p)}{p^s}\right)^{-1}. \quad (12)$$

That is, $L(s, \chi)$ has an *Euler product*. We also believe the “generalized Riemann hypothesis” that all zeros ρ of $L(\rho, \chi) = 0$ in the critical strip satisfy $\text{Re}(\rho) = \frac{1}{2}$. This would imply that the number of primes up to x that are congruent to $a \pmod q$ can be estimated as

$$\pi(x; q, a) = \frac{\pi(x)}{\phi(q)} + O(\sqrt{x} \log^2(qx)). \quad (13)$$

PUP: although there was nothing wrong here, the proofreader's comment prompted Tim to suggest a change in notation here, to match that in another article. It is all OK now.

Therefore, the generalized Riemann hypothesis implies the estimate we were hoping for (formula (11)), provided that x is a little bigger than q^2 .

In what range can we prove (11) unconditionally—that is, without the help of the generalized Riemann hypothesis? Although we can more or less translate the proof of the prime number theorem over into this new setting, we find that it gives (11) only when x is very large. In fact, x has to be bigger than an exponential in a power of q , which is a lot bigger than the “ x is a little larger than q^2 ” that we obtained from the generalized Riemann hypothesis. We see a new type of problem emerging here, in which we are asking for a good starting point for the range of x for which we obtain good estimates, as a function of the modulus q ; this does not have an analogy in our exploration of the prime number theorem. By the way, even though this bound “ x is a little larger than q^2 ” is far out of reach of current methods, it still does not seem to be the best answer; calculations reveal that (11) seems to hold when x is just a little bigger than q . So even the Riemann hypothesis and its generalizations are not powerful enough to tell us the precise behavior of the distribution of primes.

Throughout the twentieth century much thought was put in to bounding the number of zeros of Dirichlet L -functions near to the 1-line. It turns out that one can make enormous improvements in the range of x for which (11) holds (to “halfway between polynomial in q and exponential in q ”) provided there are no *Siegel zeros*. These putative zeros β of $L(s, (\frac{\cdot}{q}))$ would be real numbers with $\beta > 1 - c/\sqrt{q}$; they can be shown to be extremely rare if they exist at all.

That Siegel zeros are rare is a consequence of the *Deuring-Heilbronn phenomenon*: that zeros of L -FUNCTIONS [III.49] repel each other, rather like similarly charged particles. (This phenomenon is akin to the fact that different algebraic numbers repel one another, part of the basis of the subject of Diophantine approximation.)

How big is the smallest prime congruent to $a \bmod q$ when $(a, q) = 1$? Despite the possibility of the existence of Siegel zeros, one can prove that there is always such a prime less than $q^{5.5}$ if q is sufficiently large. Obtaining a result of this type is not difficult when there are no Siegel zeros. If there are Siegel zeros, then we go back to the explicit formula, which is similar to (7) but now concerns zeros of $L(s, \chi)$. If β is a Siegel zero, then it turns out that in the explicit formula there are now two obviously large terms: $x/\phi(q)$ and $-(\frac{a}{q})x^\beta/\beta\phi(q)$. When $(\frac{a}{q}) = 1$ it appears that they might almost cancel

(since β is close to 1), but with more care we obtain

$$x - \frac{a}{q} \frac{x^\beta}{\beta} = (x - x^\beta) + x^\beta \left(1 - \frac{1}{\beta}\right) \sim x(1 - \beta) \log x.$$

This is a smaller main term than before, but it is not too hard to show that it is bigger than the contributions of all of the other zeros combined, because the Deuring-Heilbronn phenomenon implies that the Siegel zero repels those zeros, forcing them to be far to the left. When $(\frac{a}{q}) = -1$, the same two terms tell us that if $(1 - \beta) \log x$ is small, then there are twice as many primes as we would expect up to x that are congruent to $a \bmod q$.

There is a close connection between Siegel zeros and *class numbers*, which are defined and discussed in ALGEBRAIC NUMBERS [IV.1 §7]. Dirichlet’s *class number formula* states that $L(1, (\frac{\cdot}{q})) = \pi h_{-q}/\sqrt{q}$ for $q > 6$, where h_{-q} is the class number of the field $\mathbb{Q}(\sqrt{-q})$. A class number is always a positive integer, so this result immediately implies that $L(1, (\frac{\cdot}{q})) \geq \pi/\sqrt{q}$. Another consequence is that h_{-q} is small if and only if $L(1, (\frac{\cdot}{q}))$ is small. The reason this gives us information about Siegel zeros is that one can show that the derivative $L'(\sigma, (\frac{\cdot}{q}))$ is positive (and not too small) for real numbers σ close to 1. This implies that $L(1, (\frac{\cdot}{q}))$ is small if and only if $L(s, (\frac{\cdot}{q}))$ has a real zero close to 1, that is, a Siegel zero β . When $h_{-q} = 1$, the link is more direct: it can be shown that the Siegel zero β is approximately $1 - 6/(\pi\sqrt{q})$. (There are also more complicated formulas for larger values of h_{-q} .)

These connections show that getting good lower bounds on h_{-q} is equivalent to getting good bounds on the possible range for Siegel zeros. Siegel showed that for any $\varepsilon > 0$ there exists a constant $c_\varepsilon > 0$ such that $L(1, (\frac{\cdot}{q})) \geq c_\varepsilon q^{-\varepsilon}$. His proof was unsatisfactory because by its very nature one cannot give an explicit value for c_ε . Why not? Well, the proof comes in two parts. The first assumes the generalized Riemann hypothesis, in which case an explicit bound follows easily. The second obtains a lower bound *in terms of the first counterexample* to the generalized Riemann hypothesis. So if the generalized Riemann hypothesis is true but remains unproved, then Siegel’s proof cannot be exploited to give explicit bounds. This dichotomy, between what can be proved with an explicit constant and what cannot be, is seen far and wide in analytic number theory—and when it appears it usually stems from an application of Siegel’s result, and especially its consequences for the range in which the estimate (11) is valid.

A polynomial with integer coefficients cannot always take on prime values when we substitute in an integer. To see this, note that if p divides $f(m)$ then p also divides $f(m+p), f(m+2p), \dots$. However, there are some prime-rich polynomials, a famous example being the polynomial $x^2 + x + 41$, which is prime for $x = 0, 1, 2, \dots, 39$. There are almost certainly quadratic polynomials that take on more consecutive prime values, though their coefficients would have to be very large. If we ask the more restricted question of when the polynomial $x^2 + x + p$ is prime for $x = 0, 1, 2, \dots, p-2$, then the answer, given by Rabinowitch, is rather surprising: it happens if and only if $h_{-q} = 1$, where $q = 4p - 1$. Gauss did extensive calculations of class numbers and predicted that there are just nine values of q with $h_{-q} = 1$, the largest of which is $163 = 4 \times 41 - 1$. Using the Deuring-Heilbronn phenomenon researchers showed, in the 1930s, that there is at most one q with $h_{-q} = 1$ that is not already on Gauss's list; but as usual with such methods, one could not give a bound on the size of the putative extra counterexample. It was not until the 1960s that Baker and Stark proved that there was no tenth q , both proofs involving techniques far removed from those here (in fact Heegner gave what we now understand to have been a correct proof in the 1950s but he was so far ahead of his time that it was difficult for mathematicians to appreciate his arguments and to believe that all of the details were correct). In the 1980s Goldfeld, Gross, and Zagier gave the best result to date, showing that $h_{-q} \geq \frac{1}{7700} \log q$ this time using the Deuring-Heilbronn phenomenon with the zeros of yet another type of L -function to repel the zeros of $L(s, (\frac{\cdot}{q}))$.

This idea that primes are well-distributed in arithmetic progressions except for a few rare moduli was exploited by Bombieri and Vinogradov to prove that (11) holds “almost always” when x is a little bigger than q^2 (that is, in the same range that we get “always” from the generalized Riemann hypothesis). More precisely, for given large x we have that (11) holds for “almost all” q less than $\sqrt{x}/(\log x)^2$ and for all a such that $(a, q) = 1$. “Almost all” means that, out of all q less than $\sqrt{x}/(\log x)^2$, the proportion for which (11) does not hold for every a with $(a, q) = 1$ tends to 0 as $x \rightarrow \infty$. Thus, the possibility is not ruled out that there are infinitely many counterexamples. However, since this would contradict the generalized Riemann hypothesis, we do not believe that it is so.

The *Barban-Davenport-Halberstam theorem* gives a weaker result, but it is valid for the whole feasible

range: for any given large x , the estimate (11) holds for “almost all” pairs q and a such that $q \leq x/(\log x)^2$ and $(a, q) = 1$.

5 Primes in Short Intervals

Gauss's prediction referred to the primes “around” x , so it perhaps makes more sense to interpret his statement by considering the number of primes in short intervals at around x . If we believe Gauss, then we might expect the number of primes between x and $x + y$ to be about $y/\log x$. That is, in terms of the prime-counting function π , we might expect that

$$\pi(x + y) - \pi(x) \sim \frac{y}{\log x} \quad (14)$$

for $|y| \leq x/2$. However, we have to be a little careful about the range for y . For example, if $y = \frac{1}{2} \log x$, then we certainly cannot expect to have half a prime in each interval. Obviously we need y to be large enough that the prediction can be interpreted in a way that makes sense; indeed, the Gauss-Cramér model suggests that (14) should hold when $|y|$ is a little bigger than $(\log x)^2$.

If we attempt to prove (14) using the same methods we used in the proof of the prime number theorem, we find ourselves bounding differences between ρ th powers as follows:

$$\begin{aligned} \left| \frac{(x + y)^\rho - x^\rho}{\rho} \right| &= \left| \int_x^{x+y} t^{\rho-1} dt \right| \\ &\leq \int_x^{x+y} t^{\operatorname{Re}(\rho)-1} dt \\ &\leq y(x + y)^{\operatorname{Re}(\rho)-1}. \end{aligned}$$

With bounds on the density of zeros of $\zeta(s)$ well to the right of $\frac{1}{2}$, it has been shown that (14) holds for y a little bigger than $x^{7/12}$; but there is little hope, even assuming the Riemann hypothesis, that such methods will lead to a proof of (14) for intervals of length \sqrt{x} or less.

In 1949 Selberg showed that (14) is true for “almost all” x when $|y|$ is a little bigger than $(\log x)^2$, assuming the Riemann hypothesis. Once again, “almost all” means with density tending to 1, rather than “all,” and it is feasible that there are infinitely many counterexamples, though at that time it seemed highly unlikely. It therefore came as a surprise when Maier showed, in 1984, that, for any fixed $A > 0$, the estimate (14) fails for infinitely many integers x , with $y = (\log x)^A$. His ingenious proof rests on showing that the small primes do not always have as many multiples in an interval as one might expect.

Table 2 The largest known gaps between primes.

p_n	$p_{n+1} - p_n$	$\frac{p_{n+1} - p_n}{\log^2 p_n}$
113	14	0.6264
1 327	34	0.6576
31 397	72	0.6715
370 261	112	0.6812
2 010 733	148	0.7026
20 831 323	210	0.7395
25 056 082 087	456	0.7953
2 614 941 710 599	652	0.7975
19 581 334 192 423	766	0.8178
218 209 405 436 543	906	0.8311
1 693 182 318 746 371	1132	0.9206

Let $p_1 = 2 < p_2 = 3 < \dots$ be the sequence of primes. We are now interested in the size of the gaps $p_{n+1} - p_n$ between consecutive primes. Since there are about $x/\log x$ primes up to x , the average difference is $\log x$ and we might ask how often the difference between consecutive primes is about average, whether the differences can get really small, and whether the differences can get really large. The Gauss-Cramér model suggests that the proportion of n for which the gap between consecutive primes is more than λ times the average, that is $p_{n+1} - p_n > \lambda \log p_n$, is approximately $e^{-\lambda}$; and, similarly, the proportion of intervals $[x, x + \lambda \log x]$ containing exactly k primes is approximately $e^{-\lambda} \lambda^k / k!$, a suggestion which, as we shall see, is supported by other considerations. By looking at the tail of this distribution, Cramér conjectured that $\limsup_{n \rightarrow \infty} (p_{n+1} - p_n) / (\log p_n)^2 = 1$, and the evidence we have *seems* to support this (see table 2).

The Gauss-Cramér model does have a big drawback: it does not “know any arithmetic.” In particular, as we noted earlier, it does not predict divisibility by small primes. One manifestation of this failing is that it predicts that there should be just about as many gaps of length 1 between primes as there are of length 2. However, there is only one gap of length 1, since if two primes differ by 1, then one of them must be even, whereas there are many examples of pairs of primes differing by 2, and there are believed to be infinitely many. For the model to make correct conjectures about prime pairs, we must consider divisibility by small primes in the formulation of the model, which makes it rather more complicated. Since there are these glaring errors in the simpler model, Cramér’s conjecture for the largest gaps between consecutive primes

must be treated with a degree of suspicion. And in fact, if one corrects the model to account for divisibility by small primes, one is led to conjecture that $\limsup_{n \rightarrow \infty} (p_{n+1} - p_n) / (\log p_n)^2$ is greater than $\frac{9}{8}$.

Finding large gaps between primes is equivalent to finding long sequences of composite numbers. How about trying to do this explicitly? For example, we know that $n! + j$ is composite for $2 \leq j \leq n$, as it is divisible by j . Therefore we have a gap of length at least n between consecutive primes, the first of which is the largest prime less than or equal to $n! + 1$. However, this observation is not especially helpful, since the average gap between primes around $n!$ is $\log(n!)$, which is approximately equal to $n \log n$, whereas we are looking for gaps that are *larger* than the average. However, it is possible to generalize this argument and show that there are indeed long sequences of consecutive integers, each with a small prime factor. In the 1930s, Erdős reformulated the question as follows. Fix a positive integer z , and for each prime $p \leq z$ choose an integer a_p in such a way that, for as large an integer y as possible, every positive integer $n \leq y$ satisfies at least one of the congruences $n \equiv a_p \pmod{p}$. Now let X be the product of all the primes up to z (which means, by the prime number theorem, that $\log X$ is about z), and let x be the integer between X and $2X$ such that $x \equiv -a_p \pmod{p}$ for every $p \leq z$. (This integer exists, by the *Chinese remainder theorem*.) If m is an integer between $x + 1$ and $x + y$, then $m - x$ is a positive integer less than y , so $m - x \equiv a_p \pmod{p}$ for some prime $p \leq z$. Since $x \equiv -a_p \pmod{p}$, it follows that m is divisible by p . Thus, all the integers from $x + 1$ to $x + y$ are composite. Using this basic idea, it can be shown that there are infinitely many primes p_n for which $p_{n+1} - p_n$ is about $(\log p_n)(\log \log p_n)$, which is significantly larger than the average but nowhere close to Cramér’s conjecture.

6 Gaps between Primes that Are Smaller than the Average

We have just seen how to show that there are infinitely many pairs of consecutive primes whose difference is much bigger than the average: that is, $\limsup_{n \rightarrow \infty} (p_{n+1} - p_n) / (\log p_n) = \infty$. We would now like to show that there are infinitely many pairs of consecutive primes whose difference is much smaller than the average: that is, $\liminf_{n \rightarrow \infty} (p_{n+1} - p_n) / (\log p_n) = 0$. Of course, it is believed that there are infinitely many pairs of primes that differ by 2, but this question seems intractable for now.

Until recently researchers had very little success with the question of small gaps; the best result before 2000 was that there are infinitely many gaps of size less than one-quarter of the average. However, a recent method of Goldston, Pintz, and Yıldırım, which counts primes in short intervals with simple weighting functions, proves that $\liminf_{n \rightarrow \infty} (p_{n+1} - p_n) / (\log p_n) = 0$, and even that there are infinitely many pairs of consecutive primes with difference no larger than about $\sqrt{\log p_n}$. Their proof, rather surprisingly, rests on estimates for primes in arithmetic progressions; in particular, that (11) holds for almost all q up to \sqrt{x} (as discussed earlier). Moreover, they obtain a conditional result of the following kind: if in fact (11) holds for almost all q up to a little larger than \sqrt{x} , then it follows that there exists an integer B such that $p_{n+1} - p_n \leq B$ for infinitely many primes p_n .

7 Very Small Gaps between Primes

There appear to be many pairs of primes that differ by two, like 3 and 5, 5 and 7, ..., the so-called *twin primes*, though no one has yet proved that there are infinitely many. In fact, for every even integer $2k$ there seem to be many pairs of primes that differ by $2k$, but again no one has yet proved that there are infinitely many. This is one of the outstanding problems in the subject.

In a similar vein is Goldbach's conjecture from the 1760s: is it true that every even integer greater than 2 is the sum of two primes? This is still an open question, and indeed a publisher recently offered a million dollars for its solution. We know it is true for almost all integers, and it has been computer tested for every even integer up to 4×10^{14} . The most famous result on this question is due to Chen (1966), who showed that every even integer can be written as the sum of a prime and a second integer that has *at most two* prime factors (that is, it could be a prime or an "almost-prime").

In fact, GOLDBACH [VI.17] never asked this question. He asked Euler, in a letter in the 1760s, whether every integer greater than 1 can be written as the sum of at most three primes, which would imply what we now call the "Goldbach conjecture." In the 1920s Vinogradov showed that every sufficiently large odd integer can be written as the sum of three primes (and thus every sufficiently large even integer can be written as the sum of four primes). We actually believe that every odd integer greater than 5 is the sum of three primes but the known proofs only work once the numbers involved are large enough. In this case we can be explicit about "sufficiently large"—at the moment the proof needs them

to be at least e^{5700} , but it is rumored that this may soon be substantially reduced, perhaps even to 7.

To guess at the precise number of prime pairs $q, q+2$ with $q \leq x$ we proceed as follows. If we do not consider divisibility by the small primes, then the Gauss–Cramér model suggests that a random integer up to x is prime with probability roughly $1/\log x$, so we might expect $x/(\log x)^2$ prime pairs $q, q+2$ up to x . However, we do have to account for the small primes, as the $q, q+1$ example shows, so let us consider 2-divisibility. The proportion of random pairs of integers that are both odd is $\frac{1}{4}$, whereas the proportion of random q such that q and $q+2$ are both odd is $\frac{1}{2}$. Thus we should adjust our guess $x/(\log x)^2$ by a factor $(\frac{1}{2})/(\frac{1}{4}) = 2$. Similarly, the proportion of random pairs of integers that are both not divisible by 3 (or indeed by any given odd prime p) is $(\frac{2}{3})^2$ (and $(1 - 1/p)^2$, respectively), whereas the proportion of random q such that q and $q+2$ are both not divisible by 3 (or by prime p) is $\frac{1}{3}$ (and $(1 - 2/p)$, respectively). Adjusting our formula for each prime p we end up with the prediction

$$\begin{aligned} \#\{q \leq x : q \text{ and } q+2 \text{ both prime}\} \\ \sim 2 \prod_{p \text{ an odd prime}} \frac{(1 - 2/p)}{(1 - 1/p)^2} \frac{x}{(\log x)^2}. \end{aligned}$$

This is known as the *asymptotic twin-prime conjecture*. Despite its plausibility there do not seem to be any practical ideas around for turning the heuristic argument above into something rigorous. The one good unconditional result known is that the number of twin primes less than or equal to x is never more than four times the quantity we have just predicted. One can make a more precise prediction replacing $x/(\log x)^2$ by $\int_2^x (1/(\log t)^2) dt$, and then we expect that the difference between the two sides is no more than $c\sqrt{x}$ for some constant $c > 0$, a guesstimate that is well supported by computational evidence.

A similar method allows us to make predictions for the number of primes in any polynomial-type patterns. Let $f_1(t), f_2(t), \dots, f_k(t) \in \mathbb{Z}[t]$ be distinct irreducible polynomials of degree greater than or equal to 1 with positive leading coefficient, and define $\omega(p)$ to be the number of integers $n \pmod{p}$ for which p divides $f_1(n)f_2(n) \cdots f_k(n)$. (In the case of twin primes above we have $f_1(t) = t, f_2(t) = t+2$ with $\omega(2) = 1$ and $\omega(p) = 2$ for all odd primes p .) If $\omega(p) = p$ then p always divides at least one of the polynomial values, so they can be simultaneously prime just finitely often (an example of this is when $f_1(t) = t, f_2(t) = t+1$, in

which case $\omega(2) = 2$). Otherwise we have an *admissible set* of polynomials for which we predict that the number of integers n less than x for which all of $f_1(n), f_2(n), \dots, f_k(n)$ are prime is about

$$\prod_{p \text{ prime}} \frac{(1 - \omega_f(p)/p)}{(1 - 1/p)^k} \times \frac{x}{\log |f_1(x)| \log |f_2(x)| \cdots \log |f_k(x)|} \quad (15)$$

once x is sufficiently large. One can use a similar heuristic to make predictions in Goldbach's conjecture, that is, for the number of pairs of primes p, q for which $p + q = 2N$. Again, these predictions are very well matched by the computational evidence.

There are just a few cases of conjecture (15) that have been proved. Modifications of the proof of the prime number theorem give such a result for admissible polynomials $qt + a$ (in other words, for primes in arithmetic progressions) and for admissible $at^2 + btu + cu^2 \in \mathbb{Z}[t, u]$ (as well as some other polynomials in two variables of degree two). It is also known for a certain type of polynomial in n variables of degree n (the admissible "norm-forms").

There was little improvement on this situation during the twentieth century until quite recently, when, by very different methods, Friedlander and Iwaniec broke through this stalemate showing such a result for the polynomial $t^2 + u^4$, and then Heath-Brown did so for any admissible homogeneous polynomial in two variables of degree three.

Another truly extraordinary breakthrough occurred recently with a result of Green and Tao, proved in 2004, which states that for every k there are infinitely many k -term arithmetic progressions of primes: that is, pairs of integers a, d such that $a, a+d, a+2d, \dots, a+(k-1)d$ are all prime. Green and Tao are currently hard at work attempting to show that the number of k -term arithmetic progressions of primes is indeed well approximated by (15). They are also extending their results to other families of polynomials.

8 Gaps between Primes Revisited

In the 1970s Gallagher deduced from the conjectured prediction (15) (with $f_j(t) = t + a_j$) that the proportion of intervals $[x, x + \lambda \log x]$ which contain exactly k primes is close to $e^{-\lambda} \lambda^k / k!$ (as was also deduced, in section 5 above, from the Gauss-Cramér heuristics). This has recently been extended to support the prediction that, as we vary x from X to $2X$, the number of primes in the interval $[x, x + y]$ is normally distributed with

mean $\int_x^{x+y} (1/\log t) dt$ and variance $(1 - \delta)y/\log x$, where δ is some constant strictly between 0 and 1 and we take y to be x^δ .

When $y > \sqrt{x}$ the Riemann zeta function supplies information on the distribution of primes in intervals $[x, x + y]$ via the explicit formula (7). Indeed, when we compute the "variance"

$$\frac{1}{X} \int_X^{2X} \left(\sum_{\substack{p \text{ prime,} \\ x < p \leq x+y}} \log p - y \right)^2 dx$$

using the explicit formula we obtain a sum of terms of the form $\int_X^{2X} x^{i(y_j - y_k)} dx$. Here we are assuming the Riemann hypothesis and writing the zeros of $\zeta(s)$ as $\frac{1}{2} \pm iy_n$ with $0 < y_1 < y_2 < \dots$. This sum is dominated by the terms corresponding to those pairs y_j, y_k for which $|y_j - y_k|$ is small (in which case there is little cancellation in the integral). Therefore, in order to understand the variance for the distribution of primes in short intervals we need to understand the distribution of the zeros of $\zeta(s)$ in short intervals. In 1973 Montgomery investigated this and suggested that the proportion of pairs of zeros of $\zeta(s)$ whose difference is less than α times the average gap between consecutive zeros is given by the integral

$$\int_0^\alpha \left(1 - \left(\frac{\sin \pi \theta}{\pi \theta} \right)^2 \right) d\theta, \quad (16)$$

and he proved an equivalent form of this in a limited range. If the zeros were placed "randomly," then (16) would be replaced by α . In fact (16) is about $\frac{1}{9}\alpha^3$ for small α , which is far smaller than α . This means that there are far fewer pairs of zeros of $\zeta(s)$ that are close together than one might expect, which we express informally by saying that the zeros of $\zeta(s)$ *repel* one another.

In a now-famous conversation that took place at the Institute for Advanced Study in Princeton, Montgomery mentioned his ideas to the physicist Freeman Dyson. Dyson immediately recognized (16) as a function that comes up in modeling energy levels in quantum chaos. Believing that this was unlikely to be a coincidence, he suggested that the zeros of the Riemann zeta function are distributed, *in all aspects*, like energy levels, which are in turn modeled on the distribution of EIGENVALUES [I.3 §4.3] of random HERMITIAN MATRICES [III.52 §3]. There is now substantial computational and theoretical evidence that Dyson's suggestion is correct and can be extended to Dirichlet L -functions, as well as other types of L -functions, and even to other statistics about L -functions.

One note of caution. Few of the conjectured consequences of this new “random matrix theory” have been unconditionally proved, or seem likely to be in the foreseeable future. It simply provides a tool to make predictions where that was too difficult to do before. However, there is at least one key question about which we still cannot make a well-substantiated prediction: how big does $\zeta(s)$ get on the $\frac{1}{2}$ -line? One can show that $\log |\zeta(\frac{1}{2} + it)|$ gets larger than $\sqrt{\log T}$ for values of t close to T , and that it gets no larger than $\log T$. However, it is unclear, even if we do not insist on a rigorous proof, whether the true maximal order is nearer the upper or lower bound.

9 Sieve Methods

Almost all of our discussion so far has been about developments of Riemann’s approach to counting primes. This approach is very delicate and not as adaptable as one might wish to many natural questions (such as counting k -tuples of primes $n+a_1, n+a_2, \dots, n+a_k$). However, one can go back to *sieve methods*, which are modifications of the sieve of Eratosthenes, and at least get upper bounds. For example, suppose we want to find an upper bound for the number of prime pairs $n, n+2$ with $N < n \leq 2N$. One possibility would be to fix a number y and determine for how many pairs $n, n+2$ with $N < n \leq 2N$ it is the case that neither n nor $n+2$ has a prime factor less than y . If we took y to be $(2N)^{1/2}$, then this method would exactly count the twin primes, but it seems to be far too difficult to implement. But it turns out that if instead we take y to be a small power of N , then the calculations become much easier and there are ways of obtaining good bounds. (However, these bounds become less accurate as the power gets closer to $\frac{1}{2}$.)

In the 1920s Brun showed how to make the principle of inclusion–exclusion into a useful tool in this type of question. This principle is best exhibited when counting the number of integers n in a set S that are coprime to given integer m . We begin with the number of integers in S , which is obviously more than the quantity we seek. Next, we subtract, for each prime p dividing m , the number of integers in S that are divisible by p . If $n \in S$ is divisible by exactly r prime factors of m , then we have counted $1 + r \times (-1)$ for the contribution of n so far, which is less than or equal to 0, and less than 0 for $r \geq 2$; whereas we wanted to count 0 when $r \geq 2$ (since n is not coprime to m). Thus we obtain a number that is less than the quantity we seek. To compensate

for that, we add back in the number of integers in S divisible by pq for each pair of primes $p < q$ which divide m . We have now counted $1 + r \times (-1) + \binom{r}{2} \times 1$ for the contribution of n , which is greater than or equal to 0, and greater than 0 for $r \geq 3$. Similarly, we subtract the number of integers divisible by pqr , etc.

For each $n \in S$ we end up counting $(1 - 1)^r$ for n , where r is the number of distinct prime factors of (m, n) . Expanding this sum with the binomial theorem we may reexpress this identity as follows. Let $\chi_m(n) = 1$ if $(n, m) = 1$ and 0 otherwise. Then

$$\chi_m(n) = \sum_{d|(m,n)} \mu(d),$$

where $\mu(m)$, the Möbius function, equals 0 if m is divisible by the square of a prime and equals $(-1)^{\omega(m)}$ otherwise, where $\omega(m)$ is the number of distinct prime factors of m .

The inclusion–exclusion inequalities just discussed may be obtained from

$$\sum_{\substack{d|(m,n) \\ \omega(d) \leq 2k+1}} \mu(d) \leq \chi_m(n) \leq \sum_{\substack{d|(m,n) \\ \omega(d) \leq 2k}} \mu(d),$$

which holds for any $k \geq 0$, by summing over all $n \in S$.

The reason for using these abbreviated sums rather than the complete sum is that there are far fewer terms and thus, when one sums over values of n , there will be far fewer rounding errors (remember that it was rounding errors that sank our attempt to estimate the number of primes up to x using the sieve of Eratosthenes). On the other hand, they have the disadvantage that they cannot possibly give the exact answer, since they are missing many appropriate terms. However, with a judicious choice of k the missing terms do not contribute much to the complete sum and we get a good answer.

Minor variants work well for many questions. In the “combinatorial sieve” one selects which d are part of the upper and lower bound sums, not by counting the total number of prime factors they contain but instead using other criteria, such as the numbers of prime factors of d in each of several intervals. Using such a method, Brun showed that there cannot be too many twin primes $p, p+2$; indeed, the sum of $1/p$, over all primes p for which $p+2$ is also prime, converges, in contrast with (3).

In the “Selberg upper bound sieve” one comes up with some numbers λ_d that are nonzero only when $d \leq D$ (where D is chosen to be not too large), with the property that

$$\chi_m(n) \leq \left(\sum_{d|n} \lambda_d \right)^2 \quad \text{for all } n.$$

Summing over the appropriate n one then finds the optimal solution by minimizing the resulting quadratic form. Lower bounds can also be obtained out of Selberg's methods. It was by using such methods that Chen was able to prove there are infinitely many primes p for which $p + 2$ has at most two prime factors, and that Goldston, Pintz, and Yıldırım were able to establish that there are sometimes short gaps between primes. These methods are also an essential ingredient in the work of Green and Tao. One can also get good upper bounds on the number of primes in arithmetic progressions and short intervals:

- there are never more than $2y/\log y$ primes in any interval of length y ;
- there are never more than $2x/\phi(q)\log(x/q)$ primes up to x in an arithmetic progression mod q .

Notice that in each case the log in the denominator is of the number of integers being considered (y and x/q , respectively), not $\log x$ as expected, though this will only make a significant difference if the number of integers being considered is small. Otherwise these inequalities are bigger than the expected quantity by a factor of 2. Can this “2” be improved? It will be difficult because we showed earlier that if there are Siegel zeros then we get twice as many primes as expected in certain arithmetic progressions. Therefore, if we can improve the “2” in these two formulas, then we can deduce that there are no Siegel zeros!

10 Smooth Numbers

An integer is y -smooth if all of its prime factors are less than or equal to y . A proportion $1 - \log 2$ of the integers up to x are \sqrt{x} -smooth, and indeed, for any fixed $u > 1$ there exists some number $\rho(u) > 0$ such that if $x = y^u$, then a proportion $\rho(u)$ of the integers up to x are y -smooth. This proportion does not seem to have any easy definition in general. For $1 \leq u \leq 2$ we have $\rho(u) = 1 - \log u$, but for larger u it is best defined as

$$\rho(u) = \frac{1}{u} \int_0^1 \rho(u-t) dt,$$

an *integral delay equation*. Such an equation is typical when we give precise estimates for questions that arise in sieve theory.

Questions about the distribution of smooth numbers arise frequently in the analysis of algorithms, and have consequently been the focus of a lot of recent research.

(See COMPUTATIONAL NUMBER THEORY [IV.3 §3] for an example of the use of smooth numbers.)

11 The Circle Method

Another method of analysis that plays a prominent role in this subject is the so-called *circle method*, which goes back to HARDY [VI.73] and LITTLEWOOD [VI.79]. This method uses the fact that, for any integer n ,

$$\int_0^1 e^{2i\pi nt} dt = \begin{cases} 1 & \text{if } n = 0, \\ 0 & \text{otherwise.} \end{cases}$$

For example, if we wish to count the number, $r(n)$, of solutions to the equation $p + q = n$ with p and q prime, we can express it as an integral as follows:

$$\begin{aligned} r(n) &= \sum_{\substack{p, q \leq n \\ \text{both prime}}} \int_0^1 e^{2i\pi(p+q-n)t} dt \\ &= \int_0^1 e^{-2i\pi nt} \left(\sum_{p \text{ prime}, p \leq n} e^{2i\pi pt} \right)^2 dt. \end{aligned}$$

The first equality holds because the integrand is 0 when $p + q \neq n$ and 1 otherwise, and the second is easy to check.

At first sight it looks more difficult to estimate the integral than it is to estimate $r(n)$ directly, but this is not the case. For instance, the prime number theorem for arithmetic progressions allows us to estimate $P(t) = \sum_{p \leq n} e^{2i\pi pt}$ when t is a rational ℓ/m with m small. For in this case,

$$\begin{aligned} P\left(\frac{\ell}{m}\right) &= \sum_{(a,m)=1} e^{2i\pi a\ell/m} \sum_{\substack{p \leq n, \\ p \equiv a \pmod{m}}} 1 \\ &\approx \sum_{(a,m)=1} e^{2i\pi a\ell/m} \frac{\pi(n)}{\phi(m)} = \mu(m) \frac{\pi(n)}{\phi(m)}. \end{aligned}$$

If t is sufficiently close to ℓ/m , then $P(t) \approx P(\ell/m)$; such values of t are called the *major arcs* and we believe that the integral over the major arcs gives, in total, a very good approximation to $r(n)$; indeed, we get something very close to the quantity one predicts from something like (15). Thus to prove the Goldbach conjecture we need to show that the contribution to the integral from the other values of t (that is, from the *minor arcs*) is small. In many problems one can successfully do this, but no one has yet succeeded in doing so for the Goldbach problem. Also useful is the “discrete analogue” of the above: using the identity

$$\frac{1}{m} \sum_{j=0}^{m-1} e^{2i\pi jn/m} dt = \begin{cases} 1 & \text{if } n \equiv 0 \pmod{m}, \\ 0 & \text{otherwise} \end{cases}$$

(which holds for any given integer $m \geq 1$), we have that

$$\begin{aligned} r(n) &= \sum_{\substack{p, q \leq n \\ \text{both prime}}} \frac{1}{m} \sum_{j=0}^{m-1} e^{2i\pi j(p+q-n)/m} \\ &= \sum_{j=0}^{m-1} e^{-2i\pi jn/m} P(j/m)^2 \end{aligned}$$

provided $m > n$. A similar analysis can be used here but working mod m sometimes has advantages, as it allows us to use properties of the multiplicative group mod m .

Sums like $P(j/m)$ in the paragraph above or more simple sums like $\sum_{n \leq N} e^{2i\pi n^k/m}$ are called *exponential sums*. They play a central role in many of the calculations one does in analytic number theory. There are several techniques for investigating them.

(1) It is easy to sum the geometric progression $\sum_{n \leq N} e^{2i\pi n^2/m}$. With higher-degree polynomials one can often reduce to this case; for example, by writing $n_1 - n_2 = h$ we have

$$\begin{aligned} \left| \sum_{n \leq N} e^{2i\pi n^2/m} \right|^2 &= \sum_{n_1, n_2 \leq N} e^{2i\pi(n_1^2 - n_2^2)/m} \\ &= \sum_{|h| \leq N} e^{2i\pi h^2/m} \sum_{\substack{\max\{0, -h\} < n_2 \\ \leq \min\{N, N-h\}}} e^{4i\pi h n_2/m}, \end{aligned}$$

and the inner sum is now a geometric progression.

(2) The work of Weil and Deligne, which gives very accurate results on the number of solutions to equations mod p , is ideally suited to many applications in analytic number theory. For example, the “Kloosterman sum” $\sum_{a_1 a_2 \cdots a_k \equiv b \pmod{p}} e^{2i\pi(a_1 + a_2 + \cdots + a_k)/p}$, where the a_i run over the integers mod p and $(b, p) = 1$, appears naturally in many questions; Deligne showed that it has absolute value less than or equal to $k p^{(k-1)/2}$, an extraordinary amount of cancellation in this sum which has about p^{k-1} summands, each of absolute value 1. (See THE WEIL CONJECTURES [V.38].)

(3) We discussed earlier the fact that the values of $\zeta(s)$ satisfy a symmetry about the line $\operatorname{Re}(s) = \frac{1}{2}$, given by the “functional equation.” There are other functions (called “modular functions”) that also have symmetries in the complex plane; typically the value of the function at s is related to the value of the function at $(\alpha s + \beta)/(\gamma s + \delta)$, for some integers $\alpha, \beta, \gamma, \delta$ satisfying $\alpha\delta - \beta\gamma = 1$. Sometimes an exponential sum can be

related to the value of a modular function, and subsequently to the value of that modular function at another point, using the symmetry of the function.

12 More L-Functions

There are many types of L -functions beyond Dirichlet L -functions, some of which are well understood, some not. The type that has received the most attention recently is a class of L -functions that can be associated with elliptic curves (see ARITHMETIC GEOMETRY [IV.5 §5.1]). An *elliptic curve* E is given by an equation of the form $y^2 = x^3 + ax + b$, where the *discriminant* $4a^3 + 27b^2$ is nonzero. The associated L -function $L(E, s)$ is most easily described in terms of its Euler product:

$$L(E, s) = \prod_p \left(1 - \frac{a_p}{p^s} + \frac{p}{p^{2s}} \right)^{-1}. \quad (17)$$

Here a_p is an integer which, for primes p not dividing $4a^3 + 27b^2$, is defined to be p minus the number of solutions $(x, y) \pmod{p}$ to the equation $y^2 \equiv x^3 + ax + b \pmod{p}$. It can be shown that each $|a_p|$ is less than $2\sqrt{p}$, so the Euler product above converges absolutely when $\operatorname{Re}(s) > \frac{3}{2}$. Therefore, (17) is a good definition for these values of s . Can we now extend it to the whole of the complex plane, as we did for $\zeta(s)$? This is a very deep problem—the answer is yes; in fact, it is the celebrated theorem of Andrew Wiles that implied FERMAT’S LAST THEOREM [V.12].

Another interesting question is to understand the distribution of values of $a_p/2\sqrt{p}$ as we range over primes p . These all lie in the interval $[-1, 1]$. One might expect them to be uniformly distributed in the interval, but in fact this is never the case. As discussed in ALGEBRAIC NUMBERS [IV.1] one can write $a_p = \alpha_p + \bar{\alpha}_p$, where $|\alpha_p| = \sqrt{p}$, and α_p is called the Weil number. If we write $\alpha = \sqrt{p}e^{i\theta_p}$, then $a_p = 2\sqrt{p} \cos(\theta_p)$ for some angle $\theta_p \in [0, \pi]$. We can then think of θ_p as belonging to the upper half of a circle. The surprise is that for almost all elliptic curves the θ_p are not uniformly distributed, which would mean the proportion in a certain arc would be proportional to the length of that arc. Rather, they are distributed in such a way that the proportion of them in any given arc is proportional to the area under that arc. This is a recent result of Richard Taylor.

The correct analogue of the Riemann hypothesis for $L(E, s)$ turns out to be that all the nontrivial zeros lie on the line $\operatorname{Re}(s) = 1$. This is believed to be true. Moreover, it is believed that they, like the zeros of $\zeta(s)$,

are distributed according to the rules that govern the eigenvalues of randomly chosen matrices.

These L -functions often have zeros at $s = 1$ (which is linked to THE BIRCH-SWINNERTON-DYER CONJECTURE [V.4]) and these zeros repel zeros of Dirichlet L -functions (which is what was used by Goldfeld, Gross, and Zagier, as mentioned in section 4, to get their lower bound on h_{-q}).

L -functions arise in many areas of arithmetic geometry, and their coefficients typically describe the number of points satisfying certain equations mod p . The *Langlands program* seeks to understand these connections at a deep level.

It seems that every “natural” L -function has many of the same analytic properties as those discussed in this article. Selberg has proposed that this phenomenon should be even more general. Consider sums $A(s) = \sum_{n \geq 1} a_n/n^s$ that

- are well-defined when $\operatorname{Re}(s) > 1$,
- have an Euler product $\prod_p (1 + b_p/p^s + b_{p^2}/p^{2s} + \dots)$ in this (or an even smaller) region,
- have coefficients a_n that are smaller than any given power of n , once n is sufficiently large,
- satisfy $|b_n| < \kappa n^\theta$ for some constants $\theta < \frac{1}{2}$ and $\kappa > 0$.

Selberg conjectures that we should be able to give a good definition to $A(s)$ on the whole complex plane, and that $A(s)$ should have a symmetry connecting the value of $A(s)$ with $A(1-s)$. Furthermore, he conjectures that the Riemann hypothesis should hold for $A(s)$!

The current wishful thinking is that Selberg’s family of L -functions is precisely the same as those considered by Langlands.

13 Conclusion

In this article we have described current thinking on several key questions about the distribution of primes. It is frustrating that after centuries of research so little has been proved, the primes guarding their mysteries so jealously. Each new breakthrough seems to require brilliant ideas and extraordinary technical prowess. As EULER [VI.19] wrote in 1770:

Mathematicians have tried in vain to discover some order in the sequence of prime numbers but we have every reason to believe that there are some mysteries which the human mind will never penetrate.

Further Reading

Hardy and Wright’s classic book (1980) stands alone among introductory number theory texts for the quality of its discussion of analytic topics. The best introduction to the heart of analytic number theory is the masterful book by Davenport (2000). Everything you have ever wanted to know about the Riemann zeta function is in Titchmarsh (1986). Finally, there are two recently released books by modern masters of the subject (Iwaniec and Kowalski 2004; Montgomery and Vaughan 2006) that introduce the reader to the key issues of the subject.

The reference list below includes several papers, significant for this article, whose content is not discussed in any of the listed books.

- Davenport, H. 2000. *Multiplicative Number Theory*, 3rd edn. New York: Springer.
- Deligne, P. 1977. Applications de la formule des traces aux sommes trigonométriques. In *Cohomologie Étale* (SGA 4 1/2). Lecture Notes in Mathematics, volume 569. New York: Springer.
- Green, B., and T. Tao. 2008. The primes contain arbitrarily long arithmetic progressions. *Annals of Mathematics* 167: 481–547.
- Hardy, G. H., and E. M. Wright. 1980. *An Introduction to the Theory of Numbers*, 5th edn. Oxford: Oxford Science Publications.
- Ingham, A. E. 1949. Review 10,595c (MR0029411). *Mathematical Reviews*. Providence, RI: American Mathematical Society.
- Iwaniec, H., and E. Kowalski. 2004. *Analytic Number Theory*. Colloquium Publications, volume 53. Providence, RI: American Mathematical Society.
- Montgomery, H. L., and R. C. Vaughan. 2006. *Multiplicative Number Theory I: Classical Theory*. Cambridge: Cambridge University Press.
- Soundararajan, K. 2007. Small gaps between prime numbers: the work of Goldston–Pintz–Yıldırım. *Bulletin of the American Mathematical Society* 44:1–18.
- Titchmarsh, E. C. 1986. *The Theory of the Riemann Zeta-Function*, 2nd edn. Oxford: Oxford University Press.

IV.3 Computational Number Theory

Carl Pomerance

1 Introduction

Historically, computation has been a driving force in the development of mathematics. To help measure the sizes of their fields, the Egyptians invented geometry. To help predict the positions of the planets, the Greeks invented trigonometry. Algebra was invented to deal

with equations that arose when mathematics was used to model the world. The list goes on, and it is not just historical. If anything, computation is more important than ever. Much of modern technology rests on algorithms that compute quickly: examples range from the WAVELETS [VII.3] that allow CAT scans, to the numerical extrapolation of extremely complex systems in order to predict weather and global warming, and to the combinatorial algorithms that lie behind Internet search engines (see THE MATHEMATICS OF ALGORITHM DESIGN [VII.5 §6]).

In pure mathematics we also compute, and many of our great theorems and conjectures are, at root, motivated by computational experience. It is said that GAUSS [VI.26], who was an excellent computationalist, needed only to work out a concrete example or two to discover, and then prove, the underlying theorem. While some branches of pure mathematics have perhaps lost contact with their computational origins, the advent of cheap computational power and convenient mathematical software has helped to reverse this trend.

One mathematical area where the new emphasis on computation can be clearly felt is number theory, and that is the main topic of this article. A prescient call-to-arms was issued by Gauss as long ago as 1801:

The problem of distinguishing prime numbers from composite numbers, and of resolving the latter into their prime factors, is known to be one of the most important and useful in arithmetic. It has engaged the industry and wisdom of ancient and modern geometers to such an extent that it would be superfluous to discuss the problem at length. Nevertheless we must confess that all methods that have been proposed thus far are either restricted to very special cases or are so laborious and difficult that even for numbers that do not exceed the limits of tables constructed by estimable men, they try the patience of even the practiced calculator. And these methods do not apply at all to larger numbers.... Further, the dignity of the science itself seems to require that every possible means be explored for the solution of a problem so elegant and so celebrated.

Factorization into primes is a very basic issue in number theory, but essentially all branches of number theory have a computational component. And in some areas there is such a robust computational literature that we discuss the algorithms involved as mathematically interesting objects in their own right. In this article we will briefly present a few examples of the computational spirit: in analytic number theory (the distribution of primes and the Riemann hypothesis);

in Diophantine equations (Fermat's last theorem and the ABC conjecture); and in elementary number theory (primality and factorization). A secondary theme that we shall explore is the strong and constructive interplay between computation, heuristic reasoning, and conjecture.

2 Distinguishing Prime Numbers from Composite Numbers

The problem is simple to state. Given an integer $n > 1$, decide if n is prime or composite. And we all know an algorithm. Divide n by each positive integer in turn. Either we find a proper factor, in which case we know that n is composite, or we do not, in which case we know that n is prime. For example, take $n = 269$. It is odd, so it has no even divisors. It is not a multiple of 3, so it has no divisor which is a multiple of 3. Continuing, we rule out 5, 7, 11, and 13. The next possibility, 17, has a square that is greater than 269, which means that if 269 were a multiple of 17, then it would also have to be a multiple of some number less than 17. Since we have ruled that out, we can stop our trial division at 13 and conclude that 269 is prime. (If we were actually carrying out the algorithm, we might try dividing 269 by 17, in which case we would discover that $269 = 15 \times 17 + 14$. At that point we would notice that the quotient, 15, is less than 17, which is what would tell us that 17^2 was greater than 269. Then we could stop.) In general, since a composite number n has a proper factor d with $d \leq \sqrt{n}$, one can give up on the trial dividing once one passes \sqrt{n} , at which point we know that n is prime.

This straightforward method is excellent for mental computation with small numbers, and for machine computation for somewhat larger numbers. But it scales poorly, in that if you double the number of digits of n , then the time for the worst case is squared; it is therefore an "exponential time" algorithm. One might tolerate such an algorithm for twenty-digit inputs, but think how long it would take to establish the primality of a forty-digit number! And you can forget about numbers with hundreds or thousands of digits. The issue of how the running time of an algorithm scales when one goes to larger inputs is absolutely paramount in measuring one algorithm against another. In contrast to the exponential time it takes to use trial division to recognize primes, consider the problem of multiplying two numbers. The school method of multiplication is to take each digit of one number in turn and multiply it by the other number, forming a parallelogram array.

One then performs an addition to obtain the answer. If you now double the number of digits in each number, then the parallelogram becomes twice as large in each dimension, so the running time grows by a factor of about 4. Multiplication of two numbers is an example of a “polynomial time” algorithm; its running time scales by a constant factor when the input length is doubled.

One might then rephrase Gauss’s call to arms as follows. Is there a polynomial time algorithm that distinguishes prime numbers from composite numbers? Is there a polynomial time algorithm that can produce a nontrivial factor of a composite number? It might not be apparent at this point that these are two different questions, since trial division does both. We will see, though, that it is convenient to separate them, as did Gauss.

Let us focus on recognizing primes. What we would like is a simply computed criterion that primes satisfy and composites do not, or vice versa. An old theorem of Wilson might just fit the bill. Note that $6! = 720$, which is just one less than a multiple of 7. Wilson’s theorem asserts that if n is prime, then $(n - 1)! \equiv -1 \pmod{n}$. (The meaning of this and similar statements is explained in MODULAR ARITHMETIC [III.60].) This cannot hold when n is composite, for if p is a prime factor of n and is smaller than n , then it is a factor of $(n - 1)!$, so it cannot possibly be a factor of $(n - 1)! + 1$. Thus, we have an ironclad criterion for primality. However, the Wilson criterion does not meet the standard of being simply computed, since we know no especially rapid way of computing factorials modulo another number. For example, Wilson predicts that $268! \equiv -1 \pmod{269}$, as we have already seen that 269 is prime. But if we did not know this already, how in the world could we quickly find the remainder when $268!$ is divided by 269? We can work out the product $268!$ one factor at a time, but this would take many more steps than trying divisors up to 17. It is hard to prove that something *cannot* be done, and in fact there is no theorem that says we cannot compute $a! \bmod b$ in polynomial time. We do know some ways of speeding up the computation over the totally naive method, but all methods known so far take exponential time. So, Wilson’s theorem initially seems promising, but in fact it is no help at all unless we can find a fast way to compute $a! \bmod b$.

How about FERMAT’S LITTLE THEOREM [III.60]? Note that $2^7 = 128$, which is 2 more than a multiple of 7. Or take $3^5 = 243$, which is 3 mod 5. Fermat’s little theorem

tells us that if n is prime and a is any integer, then $a^n \equiv a \pmod{n}$. If computing a large factorial modulo n is hard, perhaps it is also hard to compute a large power modulo n .

It cannot hurt to try it out for some moderate example to see if any ideas pop up. Take $a = 2$ and $n = 91$, so that we are trying to compute $2^{91} \bmod 91$. A powerful idea in mathematics is that of reduction. Can we reduce this computational problem to a smaller one? Notice that if we had already computed $2^{45} \bmod 91$, obtaining a remainder r_1 , say, then $2^{91} \equiv 2r_1^2 \pmod{91}$. That is, it is just a short additional calculation to get to our goal, yet the power 45 is only half as big. How to continue is clear: we further reduce to the exponent 22, which is less than half of 45. If $2^{22} \bmod 91 = r_2$, then $2^{45} \equiv 2r_2^2 \pmod{91}$. And of course 2^{22} is the square of 2^{11} , and so on. It is not so hard to “automate” this procedure: the exponent sequence

$$1, 2, 5, 11, 22, 45, 91$$

can be read directly from the binary (base 2) representation of 91 as 1011011, since the above sequence in binary is

$$1, 10, 101, 1011, 10110, 101101, 1011011.$$

These are the initial strings from the left of 1011011. And it is plain that the transition from one term to the next is either the double or the double plus 1.

This procedure scales nicely. When the number of digits of n is doubled, so is the sequence of exponents, and the time it takes to get from one exponent to the next, being a modular multiplication, is multiplied by 4. (As with naive multiplication, naive divide-with-remainder also takes four times as long when the size of the problem is doubled.) Thus, the overall time is multiplied by 8, yielding a polynomial time method. We call this the “powermod” algorithm.

So, let us try to illustrate Fermat’s little theorem, taking $a = 2$ and $n = 91$. Our sequence of powers is

$$\begin{aligned} 2^1 &\equiv 2, & 2^2 &\equiv 4, & 2^5 &\equiv 32, & 2^{11} &\equiv 46, \\ 2^{22} &\equiv 23, & 2^{45} &\equiv 57, & 2^{91} &\equiv 37, \end{aligned}$$

where each congruence is modulo 91, and each term in the sequence is found by squaring the prior one mod 91 or squaring and multiplying by 2 mod 91.

Wait a second: does Fermat’s little theorem not say that we are supposed to get 2 for the final residue? Well, yes, but this is guaranteed only if n is prime. And as you have probably already noticed, 91 is composite. In fact, the computation proves this.

Quite remarkably, here is an example of a computation that proves that n is composite, yet it does not reveal any nontrivial factorization!

You are invited to try out the powermod algorithm as above, but to change the base of the power from 2 to 3. The answer you should come to is that $3^{91} \equiv 3 \pmod{91}$: that is, the congruence for Fermat's little theorem holds. Since you already know that 91 is composite, I am sure you would not jump to the false conclusion that it is prime! So, as it stands, Fermat's little theorem can sometimes be used to recognize composites, but it cannot be used to recognize primes.

There are two interesting further points to be made regarding Fermat's little theorem. First, on the negative side, there are some composites, such as $n = 561$, where the Fermat congruence holds for *every* integer a . These numbers n are called *Carmichael numbers*, and unfortunately (from the point of view of testing primality) there are infinitely many of them, a result due to Alford, Granville, and me. But, on the positive side, if one were to choose randomly among all pairs a, n for which $a^n \equiv a \pmod{n}$, with $a < n$ and n bounded by a large number x , almost certainly (as x grows) you would choose a pair with n prime, a result of Erdős and myself.

It is possible to combine Fermat's little theorem with another elementary property of (odd) prime numbers. If n is an odd prime, there are exactly two solutions to the congruence $x^2 \equiv 1 \pmod{n}$, namely ± 1 . Actually, some composites have this property as well, but composites divisible by two different odd primes do not.

Now let us suppose that n is an odd number and that we wish to determine whether it is prime. Suppose that we pick some number a with $1 \leq a \leq n-1$ and discover that $a^{n-1} \equiv 1 \pmod{n}$. If we set $x = a^{(n-1)/2}$, then $x^2 = a^{n-1} \equiv 1 \pmod{n}$; so, by the simple property of primes just mentioned, if n is prime, then x must be ± 1 . Therefore, if we calculate $a^{(n-1)/2}$ and discover that it is not congruent to $\pm 1 \pmod{n}$, then n must be composite.

Let us try this idea with $a = 2$, $n = 561$. We know already that $2^{560} \equiv 1 \pmod{561}$, so what is $2^{280} \pmod{561}$? This too turns out to be 1, so we have not shown that 561 is composite. However, we can go further, since now we know that 2^{140} is also a square root of 1 and computing this we find that $2^{140} \equiv 67 \pmod{561}$. So now we have found a square root of 1 that is not ± 1 , which proves that 561 is composite. (Of

course, for this particular number, it is obviously divisible by 3, so there was not really any mystery about whether it was prime or composite. But the method can be used in much less obvious cases.) In practice, there is no need to backtrack from a higher exponent to a smaller one. Indeed, in order to calculate $2^{560} \pmod{561}$ by the efficient method outlined earlier, one calculates the numbers 2^{140} and 2^{280} along the way, so that this generalization of the earlier test is both quicker and stronger.

Here is the general principle that we have illustrated. Suppose that n is an odd prime and let a be an integer not divisible by n . Write $n-1 = 2^s t$, where t is odd. Then

$$\text{either } a^t \equiv 1 \pmod{n} \quad \text{or} \quad a^{2^i t} \equiv -1 \pmod{n}$$

for some $i = 0, 1, \dots, s-1$. Call this the *strong Fermat congruence*. The wonderful thing here is that, as proved independently by Monier and Rabin, there is no analogue of a Carmichael number. They showed that if n is an odd composite, then the strong Fermat congruence fails for at least three quarters of the choices for a with $1 \leq a \leq n-1$.

If you want only to be able to distinguish between primes and composites in practice, and you do not insist on proof, then you have read enough. Namely, given a large odd number n , choose twenty values of a at random from $[1, n-1]$, and begin trying to verify the strong Fermat congruence with these bases a . If it should ever fail, you may stop: the number n must be composite. And if the strong Fermat congruence holds, we might surmise that n is actually prime. Indeed, if n were composite, the Monier-Rabin theorem says that the chance that the strong Fermat congruence would hold for twenty random bases is at most 4^{-20} , which is less than one chance in a trillion. Thus we have a remarkable *probabilistic* test for primality. If it tells us that n is composite, then we know for sure that n is composite; if it tells us that n is prime, then the chances that n is not prime are so small as to be more or less negligible.

If three quarters of the numbers a in $[1, n-1]$ provide the key to an easily checkable proof that the odd composite number n is indeed composite, surely it should not be so hard to find just one! How about checking small numbers a , in order, until one is found? Excellent, but when do we stop? Let us think about this for a moment. We have given up the power of randomness and are forcing ourselves to choose sequentially among small numbers for the trial bases a . Can we

argue heuristically that they continue to behave as if they were random choices? Well, there *are* some connections among them. For example, if taking $a = 2$ does not result in a proof that n is composite, then neither will taking any power of 2. It is theoretically possible for 2 and 3 not to give proofs that n is composite but for 6 to work just fine, but this turns out not to be very common. So let us amend the heuristic and assume that we have independence for *prime* values of a . Up to $\log n \log \log n$ there are about $\log n$ primes (via the PRIME NUMBER THEOREM [V.29] discussed later in this article); so, heuristically, the probability that n is composite, but that none of these primes help us to prove it, is about $4^{-\log n} < n^{-4/3}$. Since the infinite sum $\sum n^{-4/3}$ converges, perhaps a stopping point of $\log n \log \log n$ is sufficient, at least for large n .

Miller was able to prove the slightly weaker result that a stopping point of $c(\log n)^2$ is adequate, but his proof assumes a generalization of THE RIEMANN HYPOTHESIS [V.29]. (We discuss the Riemann hypothesis below; the generalization that Miller assumes is beyond the scope of this article.) In further work, Bach was able to show that we may take $c = 2$ in this last result. Summarizing, if this generalized Riemann hypothesis holds, and if the strong Fermat congruence holds for every positive integer $a \leq 2(\log n)^2$, then n is prime. So, provided that a famous unproved hypothesis in another field of mathematics is correct, one can decide in polynomial time, via a deterministic algorithm, whether n is prime or composite. (It has been tempting to use this conditional test, for if it should ever lie to you and tell you that a particular composite number is prime, then this failure—if you were able to detect it—would be a disproof of one of the most famous conjectures in mathematics. Perhaps this is not too disastrous a failure!)

After Miller's test in the 1970s, the question continually challenging us was whether it is possible to test for primality in polynomial time without assuming unproved hypotheses. Recently, Agrawal et al. (2004) answered this question with a resounding yes. Their idea begins with a combination of the binomial theorem and Fermat's little theorem. Given an integer a , consider the polynomial $(x + a)^n$ and expand it in the usual way through the binomial theorem. Each intermediate term between the leading x^n and the trailing a^n has the coefficient $n!/(j!(n-j)!)$ for some j between 1 and $n-1$. If n is prime, then this coefficient, which is an integer, is divisible by n because n appears as a factor in the numerator that is not canceled by any

factors in the denominator. That is, the coefficient is $0 \pmod{n}$. For example, $(x + 1)^7$ is equal to

$$x^7 + 7x^6 + 21x^5 + 35x^4 + 35x^3 + 21x^2 + 7x + 1,$$

and we see each internal coefficient is a multiple of 7. Thus, we have $(x + 1)^7 \equiv x^7 + 1 \pmod{7}$. (Two polynomials are congruent mod n if corresponding coefficients are congruent mod n .) In general, if n is prime and a is any integer, then via this binomial-theorem idea and Fermat's little theorem we have

$$(x + a)^n \equiv x^n + a^n \equiv x^n + a \pmod{n}.$$

It is an easy exercise to show that this congruence in the simple case $a = 1$ is actually equivalent to primality. But as with the Wilson criterion we know no way of quickly verifying that all these coefficients are indeed divisible by n .

However, one can do more with polynomials than raise them to powers. We can also divide one polynomial by another to find a quotient and a remainder, just as we do with integers. It makes sense, for example, to say that $g(x) \equiv h(x) \pmod{f(x)}$, meaning that $g(x)$ and $h(x)$ leave the same remainder when divided by $f(x)$. We will write $g(x) \equiv h(x) \pmod{n, f(x)}$ if the remainders upon division by $f(x)$ are congruent mod n . As with the powermod algorithm for integer congruences, we can quickly compute $g(x)^n \pmod{n, f(x)}$, provided the degree of $f(x)$ is not too big. This is exactly what Agrawal et al. propose. They have an auxiliary polynomial $f(x)$ of not-too-high degree such that, if

$$(x + a)^n \equiv x^n + a \pmod{n, f(x)}$$

for each $a = 1, 2, \dots, B$, for a not-too-high bound B , then n must be in a set that contains the primes and certain composites that are easily recognized as composites. (Not all composites are hard to recognize as such, e.g., any number with a small prime factor is easy to recognize.) These ideas put together form the primality test of Agrawal et al. To give the argument in full detail one has to specify the auxiliary polynomial $f(x)$ that is used and what the bound B is, and one has to prove rigorously that it is exactly the primes which pass the test.

Agrawal et al. (2004) show that the auxiliary polynomial $f(x)$ can be taken to be the beautifully simple $x^r - 1$, with an elementary upper bound for r of about $(\log n)^5$. Doing this leads to a time bound of about $(\log n)^{10.5}$ for the algorithm. Using a numerically ineffective tool, they bring the time bound down

to $(\log n)^{7.5}$. Recently, Lenstra and I presented a not-so-simple but numerically effective method of bringing the exponent on $\log n$ down to 6. We did this by expanding the set of polynomials used beyond those of the form $x^r - 1$: in particular we used polynomials that are related to Gauss's famous algorithm for construction of certain regular n -gons with straightedge and compass (see ALGEBRAIC NUMBERS [IV.1 §13]). It was indeed satisfying to us to bring in a famous tool of Gauss to say something about his problem of distinguishing prime numbers from composite numbers.

Are the new polynomial-time primality tests good in practice? So far, the answer is no, the competition is just too tough. For example, using the arithmetic of ELLIPTIC CURVES [III.21] we can come up with bona fide proofs of primality for huge numbers. This algorithm is conjectured to run in polynomial time but we have not even proved that it always terminates. If, at the end of the day, or in this case the end of the run, we have a legitimate proof, then perhaps we can tolerate the situation of not being sure that it would work out when we started! The method, pioneered by Atkin and Morain, has recently proved the primality of a number that has over 20 000 decimal digits, and is not of some special form such as $2^n - 1$ that makes testing for primality easier. The record for the new breed of polynomial-time tests is a measly 300 digits.

For numbers of certain special forms there are much faster primality tests. Mersenne primes comprise the most famous of these forms; these are primes that are 1 less than a power of 2. It is suspected that there are infinitely many examples, but we seem to be very far from a proof of this. Just forty-three Mersenne primes are known, the record example being $2^{30402457} - 1$, a prime with more than 9.15 million decimal digits.

For much more on primality testing, and for references to various other sources, see Crandall and Pomerance (2005).

3 Factoring Composite Numbers

Compared with what we know about testing primality, our ability to factor large numbers is still in the dark ages. In fact this imbalance between the two problems forms the bulwark for the security of electronic commerce on the Internet. (See MATHEMATICS AND CRYPTOGRAPHY [VII.7] for an account of why.) This is a very important application of mathematics, but also an odd one, and not something to brag about, since it depends on the inability of mathematicians to efficiently solve a basic problem!

Nevertheless, we do have our tricks. Part of the landscape is EUCLID'S ALGORITHM [III.22] for computing the greatest common divisor (GCD) of two numbers. One might naively think that, to find the GCD of two positive integers m and n , one should find all of their divisors and pick the largest one common to the two. But Euclid's algorithm is much more efficient: the number of arithmetic steps is bounded by the logarithm of the smaller number, so not only does it run in polynomial time, it is in fact quite speedy.

So, if we can build up a special number m that may be likely to have a nontrivial factor in common with n , we can use Euclid's algorithm to discover this factor. For example, Pollard and Strassen (independently) used this idea, together with fast subroutines for multiplication and polynomial evaluation, to enhance the trial division method discussed in the last section. Somewhat miraculously, one can take the integers up to $n^{1/2}$, break them into $n^{1/4}$ subintervals of length $n^{1/4}$, and for each subinterval calculate the GCD of n with the product of all the integers in the subinterval, spending only about $n^{1/4}$ elementary steps in total. If n is composite, then at least one GCD will be larger than 1, and then a search over the first such subinterval will locate a nontrivial factor of n . To date, this algorithm is the fastest rigorous and deterministic method of factoring that we know.

Most practical factoring algorithms are based on unproved but reasonable-seeming hypotheses about the natural numbers. Although we may not know how to prove rigorously that these methods will always produce a factorization, or do so quickly, in practice they do. This situation resembles the experimental sciences, where hypotheses are tested against experiments. Our experience with certain factoring algorithms is now so overwhelming that a scientist might claim that a physical law is involved. As mathematicians, we still search for proof, but fortunately the numbers we factor do not feel the need to wait for us.

I often mention a contest problem from my high school years: factor 8051. The trick is to notice that $8051 = 90^2 - 7^2 = (90 - 7)(90 + 7)$, from which the factorization $83 \cdot 97$ can be read off. In fact every odd composite can be factored as the difference of two squares, an idea that goes back to FERMAT [VI.12]. Indeed, if n has the nontrivial factorization ab , then let $u = \frac{1}{2}(a + b)$ and $v = \frac{1}{2}(a - b)$, so that $n = u^2 - v^2$, and $a = u + v$, $b = u - v$. This method works very well if n has a divisor very close to $n^{1/2}$, as $n = 8051$ does,

but in the worst case, the Fermat method is slower than trial division.

My quadratic sieve method (which follows work of Kraitchik, Brillhart–Morrison, and Schroeppel) tries to efficiently extend Fermat’s idea to all odd composites. For example, take $n = 1649$. We start just above $n^{1/2}$ with $j = 41$, and consider the numbers $j^2 - 1649$. As j runs, we will eventually hit a value where $j^2 - 1649$ is a square, and so be able to use Fermat’s method. Let’s try it:

$$\begin{aligned} 41^2 - 1649 &= 32, \\ 42^2 - 1649 &= 115, \\ 43^2 - 1649 &= 200, \\ &\vdots \end{aligned}$$

Well, no squares yet, which is not surprising, since the Fermat method is often very poor. But wait, do the first and third lines not multiply together to give a square? Yes they do, $32 \cdot 200 = 80^2$. So, multiplying the first and third lines, and treating them as congruences mod 1649, we have

$$(41 \cdot 43)^2 \equiv 80^2 \pmod{1649}.$$

That is, we have a pair u, v with $u^2 \equiv v^2 \pmod{1649}$. This is not quite the same as having $u^2 - v^2 = 1649$, but we do have 1649 a divisor of $u^2 - v^2 = (u - v)(u + v)$. Now maybe 1649 divides one of these factors, but if it does not, then it is split between them, and so a computation of the GCD of $u - v$ (or $u + v$) with 1649 will reveal a proper factor. Now $v = 80$ and $u = 41 \cdot 43 \equiv 114 \pmod{1649}$, and so we see instantly that $u \not\equiv \pm v \pmod{1649}$, so we are in business. The GCD of $114 - 80 = 34$ with 1649 is 17. Dividing, we see that $1649 = 17 \cdot 97$, and we are done.

Can we generalize this? In trying to factor $n = 1649$ we considered consecutive values of the quadratic polynomial $f(j) = j^2 - n$ for j starting just above \sqrt{n} , and viewed these as congruences $j^2 \equiv f(j) \pmod{n}$. Then we found a set \mathcal{M} of numbers j with $\prod_{j \in \mathcal{M}} f(j)$ equal to a square, say v^2 . We then let $u = \prod_{j \in \mathcal{M}} j$, so that $u^2 \equiv v^2 \pmod{n}$. Since $u \not\equiv \pm v \pmod{n}$, we could split n via the GCD of $u - v$ and n .

There is another lesson that we can learn from our small example with $n = 1649$. We used 32 and 200 to form our square, but we ignored 115. If we had thought about it, we might have noticed from the start that 32 and 200 were more likely to be useful than 115. The reason is that 32 and 200 are *smooth numbers* (meaning that they have only small prime factors), while 115

is not smooth, having the relatively large prime factor 23. Say you have $k + 1$ positive integers that involve in their prime factorizations only the first k primes. It is an easy theorem that some nonempty subset of these numbers has product a square. The proof has us associate with each of these numbers, which can be written in the form $p_1^{a_1} p_2^{a_2} \cdots p_k^{a_k}$, an *exponent vector* (a_1, a_2, \dots, a_k) . Since squares are detected by all even exponents, we really only care whether the exponents a_i are odd or even. Thus, we think of these vectors as having coordinates 0 and 1, and when we add them (which corresponds to multiplying the underlying numbers), we do so mod 2. Since we have $k + 1$ vectors, each with only k coordinates, an easy matrix calculation leads quickly to a nonempty subset that adds up to the 0-vector. The product of the corresponding integers is then a square.

In our toy example with $n = 1649$, the first and third numbers, which are $32 = 2^5 3^0 5^0$ and $200 = 2^3 3^0 5^2$, have exponent vectors $(5, 0, 0)$ and $(3, 0, 2)$, which reduce to $(1, 0, 0)$ and $(1, 0, 0)$, so we see that the sum of them is $(0, 0, 0)$, which indicates that we have a square. We were lucky that we could make do with just two vectors, instead of the four that the above argument shows would be sufficient.

In general with the quadratic sieve, one finds smooth numbers in the sequence $j^2 - n$, forms the exponent vectors mod 2, and then uses a matrix to find a nonempty subset which adds up to the 0-vector, which then corresponds to a set \mathcal{M} for which $\prod_{j \in \mathcal{M}} f(j)$ is a square.

In addition, the “sieve” in the quadratic sieve comes in with the search for smooth values of $f(j) = j^2 - n$. These numbers are the consecutive values of a (quadratic) polynomial, so those divisible by a given prime can be found in regular places in the sequence. For example, in our illustration, $j^2 - 1649$ is divisible by 5 precisely when $j \equiv 2$ or $3 \pmod{5}$. A sieve very much like the sieve of Eratosthenes can then be used to efficiently find the special numbers j where $j^2 - n$ is smooth. A key issue, though, is how smooth a value $f(j)$ has to be for us to decide to accept it. If we choose a smaller bound for the primes involved, we do not have to find all that many of them to use the matrix method. But such very smooth values might be very rare. If we use a larger bound for the primes involved, then smooth values of $f(j)$ may be more common, but we will need many of them. Somewhere between smaller and larger is just right! In order to make the choice, it would help to know how frequently values of

an irreducible quadratic polynomial are smooth. Unfortunately, we do not have a theorem that tells us, but we can still make a good choice by assuming that this frequency is about that for a random number of the same size, an assumption that is probably correct even if it is hard to prove.

Finally, note that if the final GCD yields only a trivial factor with n , one can continue just a bit longer and find more linear dependencies, each with a fresh chance at splitting n .

These thoughts lead us to a time bound of about

$$\exp\left(\sqrt{\log n \log \log n}\right)$$

for the quadratic sieve to factor n . Instead of being exponential in the number of digits of n , as with trial division, this is exponential in about the square root of the number of digits of n . This is certainly a huge improvement, but it is still a far cry from polynomial time.

Lenstra and I actually have a rigorous random factoring method with the same time complexity as that above for the quadratic sieve. (It is random in the sense that a coin is flipped at various junctures, and decisions on what to do next depend on the outcomes of these flips. Through this process, we expect to get a bona fide factorization within the advertised time bound.) However, the method is not so computer practical, and if you had to choose in practice between the two, then you should go with the nonrigorous quadratic sieve. A triumph for the quadratic sieve was the 1994 factorization of the 129-digit RSA cryptographic challenge first published in Martin Gardner's column in *Scientific American* in 1977.

The *number field sieve*, which is another sieve-based factoring algorithm, was discovered in the late 1980s by Pollard for integers close to powers, and later developed by Buhler, Lenstra, and me for general integers. The method is similar in spirit to the quadratic sieve, but assembles its squares from the product of certain sets of algebraic integers. The number field sieve has a conjectured time complexity of the type

$$\exp(c(\log n)^{1/3}(\log \log n)^{2/3}),$$

for a value of c slightly below 2. For composite numbers beyond 100 digits or so that have no small prime factor, it is the method of choice, with the current record being 200 decimal digits.

The sieve-based factorization methods share the property that if you use them, then all composite numbers of about the same size are equally hard to factor. For instance, factoring n will be about as difficult

if n is a product of five primes each roughly near the fifth root of n as it will be if n is a product of two primes roughly near the square root of n . This is quite unlike trial division, which is happiest when there is a small prime factor. We will now describe a famous factorization method due to Lenstra that detects small prime factors before large ones, and beyond baby cases is much superior to trial dividing. This is his *elliptic curve method*.

PUP: I can confirm that this jargon is OK here.

Just as the quadratic sieve searches for a number m with a nontrivial GCD with n , so does the elliptic curve method. But where the quadratic sieve painstakingly builds up to a successful m from many small successes, the elliptic curve method hopes to hit upon m with essentially one lucky choice.

Choosing random numbers m and testing their GCD with n can also have instant success, but you can well imagine that if n has no small prime factors, then the expected time for success would be enormous. Instead, the elliptic curve method involves considerably more cleverness.

Consider first the “ $p-1$ method” of Pollard. Suppose you have a number n you wish to factor and a certain large number k . Unbeknownst to you, n has a prime factor p with $p-1$ a divisor of k , and another prime factor q with $q-1$ not a divisor of k . You can use this imbalance to split n . First of all, by Fermat's little theorem there are many numbers u with $u^k \equiv 1 \pmod{p}$ and $u^k \not\equiv 1 \pmod{q}$. Say you have one of these, and let m be $u^k - 1$ reduced mod n . Then the GCD of m and n is a nontrivial factor of n ; it is divisible by p but not by q . Pollard suggests taking k as the least common multiple of the integers to some moderate bound so that it has many divisors and perhaps a decent chance that it is divisible by $p-1$. The best case of Pollard's method is when n has a prime factor p with $p-1$ smooth (has all small prime factors—see the quadratic sieve discussion above). But if n has no prime factors p with $p-1$ smooth, Pollard's method fares poorly.

What is going on here is that corresponding to the prime p we have the multiplicative GROUP [I.3 §2.1] of the $p-1$ nonzero residues mod p . Furthermore, when doing arithmetic mod n with numbers relatively prime to n , we are, whether we realize it or not, doing arithmetic in this group. We are exploiting the fact that u^k is the group identity mod p , but not mod q .

Lenstra had the brilliant idea of using the Pollard method in the context of elliptic curve groups. There are many elliptic curve groups associated with the prime p , and therefore many chances to hit upon one

where the number of elements is smooth. Of great importance here are theorems of Hasse and Deuring. An ELLIPTIC CURVE [III.21] mod p (for $p > 3$) can be taken as the set of solutions to the congruence $y^2 \equiv x^3 + ax + b \pmod{p}$, for given integers a, b with the property that $x^3 + ax + b$ does not have repeated roots mod p . There is one additional “point at infinity” thrown in (see below). A fairly simple addition law (but not as simple as adding coordinatewise!) makes the elliptic curve into a group, with the point at infinity as the identity (see RATIONAL POINTS ON CURVES AND THE MORDELL CONJECTURE [V.32]). Hasse, in a result later generalized by WEIL [VI.93] with his famous proof of the “Riemann hypothesis for curves,” showed us that the number of elements in the elliptic curve group always lies between $p + 1 - 2\sqrt{p}$ and $p + 1 + 2\sqrt{p}$ (see THE WEIL CONJECTURES [V.38]). And Deuring proved that every number in this range is indeed associated with some elliptic curve mod p .

Say we randomly choose integers x_1, y_1, a , and then choose b so that y_1^2 is congruent to $x_1^3 + ax_1 + b \pmod{n}$. This gives us the curve with coefficients a, b and a point $P = (x_1, y_1)$ on the curve. One can then mimic the Pollard strategy, with a number k as before with many divisors, and with the point P playing the role of u . Let kP denote the k -fold sum of P added to itself using elliptic curve addition. If kP is the point at infinity on the curve considered mod p (which it will be if the number of points on the curve is a divisor of k), but not on the curve considered mod q , then this gives us a number m whose GCD with n is divisible by p and not by q . We will have factored n .

To see where m comes from it is convenient to consider the curve projectively: we take solutions (x, y, z) of the congruence $y^2z \equiv x^3 + axz^2 + bz^3 \pmod{p}$. The triple (cx, cy, cz) when $c \neq 0$ is considered to be the same as (x, y, z) . The mysterious point at infinity is now demystified; it is just $(0, 1, 0)$. And our point P is $(x_1, y_1, 1)$. (This is the mod p version of classical PROJECTIVE GEOMETRY [I.3 §6.7].) Say we work mod n and compute the point $kP = (x_k, y_k, z_k)$. Then the candidate for the number m is just z_k . Indeed, if kP is the point at infinity mod p , then $z_k \equiv 0 \pmod{p}$, and if it is not the point at infinity mod q , then $z_k \not\equiv 0 \pmod{q}$.

When Pollard’s $p - 1$ method fails, our only recourse is to raise k or give up. With the elliptic curve method, if things do not work for our randomly chosen curve, we can pick another. Corresponding to the hidden prime p in n , we are actually picking new elliptic curve groups mod p , and so gaining a fresh chance for the number of

elements in the group to be smooth. The elliptic curve method has been quite successful in factoring numbers which have a prime factor up to about fifty decimal digits, and occasionally even somewhat larger primes have been discovered.

We conjecture that the expected time for the elliptic curve method to find the least prime factor p of n is about

$$\exp(\sqrt{2 \log p \log \log p})$$

arithmetic operations mod n . What is holding us back from proving this conjecture is not lack of knowledge about elliptic curves, but rather lack of knowledge of the distribution of smooth numbers.

For more on these and other factorization methods, the reader is referred to Crandall and Pomerance (2005).

4 The Riemann Hypothesis and the Distribution of the Primes

As a teenager looking at a modest table of primes, Gauss conjectured that their frequency decays logarithmically and that $\text{li}(x) = \int_2^x (1/\log t) dt$ should be a good approximation for $\pi(x)$, the number of primes between 1 and x . Sixty years later, RIEMANN [VI.49] showed how Gauss’s conjecture can be proved if one assumes that the Riemann zeta function $\zeta(s) = \sum_n n^{-s}$ has no zeros in the complex half-plane where the real part of s is greater than $\frac{1}{2}$. The series for $\zeta(s)$ converges only for $\text{Re } s > 1$, but it may be analytically continued to $\text{Re } s > 0$, with a simple pole at $s = 1$. (For a brief description of the process of analytic continuation, see SOME FUNDAMENTAL MATHEMATICAL DEFINITIONS [I.3 §5.6].) This continuation may be seen quite concretely via the identity $\zeta(s) = s/(s-1) - s \int_1^\infty \{x\} x^{-s-1} dx$, with $\{x\}$ the fractional part of x (so that $\{x\} = x - [x]$): note that this integral converges quite nicely in the half-plane $\text{Re } s > 0$. In fact, via Riemann’s functional equation mentioned below, $\zeta(s)$ can be continued to a meromorphic function in the whole complex plane, with the single pole at $s = 1$.

The assertion that $\zeta(s) \neq 0$ for $\text{Re } s > \frac{1}{2}$ is known as the RIEMANN HYPOTHESIS [IV.2 §3]; arguably it is the most famous unsolved problem in mathematics. Though HADAMARD [VI.65] and DE LA VALLÉE POUSSIN [VI.67] were able in 1896 to prove (independently) a weak form of Gauss’s conjecture known as the PRIME NUMBER THEOREM [V.29], the apparent breathtaking strength of the approximation $\text{li}(x)$ to $\pi(x)$ is uncanny.

For example, take $x = 10^{22}$. We have

$$\pi(10^{22}) = 201\,467\,286\,689\,315\,906\,290$$

exactly, and, to the nearest integer, we have

$$\text{li}(10^{22}) \approx 201\,467\,286\,691\,248\,261\,497.$$

As you can plainly see, Gauss's guess is right on the money!

The numerical computation of $\text{li}(x)$ is simple via numerical methods for integration, and it is directly obtainable in various mathematics computing packages. However, the computation of $\pi(10^{22})$ (due to Gourdon) is far from trivial. It would be far too laborious to count these approximately 2×10^{20} primes one by one, so how are they counted? In fact, we have various combinatorial tricks to count without listing everything. For example, one does not need to count one by one to see that there are exactly $2[10^{22}/6] + 1$ integers in the interval from 1 to 10^{22} that are relatively prime to 6. Rather, one thinks of these numbers grouped in blocks of six, with two in each block coprime to 6. (The “+1” comes from the partial block at the end.) Building on early ideas of Meissel and Lehmer, Lagarias, Miller, and Odlyzko presented an elegant combinatorial method for computing $\pi(x)$ that takes about $x^{2/3}$ elementary steps. The method was refined by Deléglise and Rivat, and then Gourdon found a way to distribute the computation to many computers.

From work of von Koch, and later Schoenfeld, we know that the Riemann hypothesis is *equivalent* to the assertion that

$$|\pi(x) - \text{li}(x)| < \sqrt{x} \log x \quad (1)$$

for all $x \geq 3$ (see Crandall and Pomerance 2005, exercise 1.37). Thus, the mammoth calculation of $\pi(10^{22})$ might be viewed as computational evidence for the Riemann hypothesis—in fact, if the count had turned out to violate (1), we would have had a disproof.

It may not be obvious what (1) has to do with the location of the zeros of $\zeta(s)$. To understand the connection, let us first dismiss the so-called “trivial” zeros, which occur at each negative even integer. The nontrivial zeros ρ are known to be infinite in number, and, as mentioned above, are conjectured to satisfy $\text{Re } \rho \leq \frac{1}{2}$. There are certain symmetries among these zeros: indeed, if ρ is a zero, then so are $\bar{\rho}$, $1 - \rho$, and $1 - \bar{\rho}$. Therefore, the Riemann hypothesis is the assertion that every nontrivial zero has real part equal to $\frac{1}{2}$. (The symmetry with ρ and $1 - \rho$, which follows from Riemann's functional

equation $\zeta(1-s) = 2(2\pi)^{-s} \cos(\frac{1}{2}\pi s) \Gamma(s) \zeta(s)$, perhaps provides some heuristic support for the Riemann hypothesis.)

The connection to prime numbers begins with THE FUNDAMENTAL THEOREM OF ARITHMETIC [V.16], which yields the identity

$$\begin{aligned} \zeta(s) &= \sum_{n=1}^{\infty} n^{-s} = \prod_{p \text{ prime}} \sum_{j=0}^{\infty} p^{-js} \\ &= \prod_{p \text{ prime}} (1 - p^{-s})^{-1}, \end{aligned}$$

a product that converges when $\text{Re } s > 1$. Thus, taking the logarithmic derivative (that is, taking the logarithm of both sides and then differentiating), we have

$$\frac{\zeta'(s)}{\zeta(s)} = - \sum_{p \text{ prime}} \frac{\log p}{p^s - 1} = - \sum_{p \text{ prime}} \sum_{j=1}^{\infty} \frac{\log p}{p^{js}}.$$

That is, if we define $\Lambda(n)$ to be $\log p$ if $n = p^j$ for a prime p and an integer $j \geq 1$, and $\Lambda(n) = 0$ if n is not of this form, then we have the identity

$$\sum_{n=1}^{\infty} \frac{\Lambda(n)}{n^s} = - \frac{\zeta'(s)}{\zeta(s)}.$$

Through various relatively routine calculations, one can then relate the function

$$\psi(x) = \sum_{n \leq x} \Lambda(n)$$

to the residues at the poles of ζ'/ζ , which correspond to the zeros (and single pole) of ζ . In fact, as Riemann showed, we have the following beautiful formula:

$$\psi(x) = x - \sum_{\rho} \frac{x^{\rho}}{\rho} - \log(2\pi) - \frac{1}{2} \log(1 - x^{-2})$$

if x itself is not a prime or prime power, and where the sum over the nontrivial zeros ρ of ζ is to be understood in the symmetric sense where we sum over those ρ with $|\text{Im } \rho| < T$ and let $T \rightarrow \infty$. Through elementary manipulations, an understanding of the function $\psi(x)$ readily gives an equivalent understanding of $\pi(x)$, and it should be clear now that $\psi(x)$ is intimately connected to the nontrivial zeros ρ of ζ .

The function $\psi(x)$ defined above has a simple interpretation. It is the logarithm of the least common multiple of the integers in the interval $[1, x]$. As with (1) we have an elementary translation of the Riemann hypothesis: it is equivalent to the assertion that

$$|\psi(x) - x| < \sqrt{x} \log^2 x$$

for all $x \geq 3$. This inequality involves only the elementary concepts of least common multiple, natural

logarithm, absolute value, and square root, yet it is equivalent to the Riemann hypothesis.

A number of nontrivial zeros ρ of $\zeta(s)$ have actually been calculated and it has been verified that they lie on the line $\text{Re } s = \frac{1}{2}$. One might wonder how someone can computationally verify that a complex number ρ has $\text{Re } \rho = \frac{1}{2}$. For example, suppose that we are carrying calculations to (an unrealistically large) 10^{10} significant digits, and suppose we come across a zero with real part $\frac{1}{2} + 10^{-10^{100}}$. It would be far beyond the precision of the calculation to be able to distinguish this number from $\frac{1}{2}$ itself. Nevertheless, we do have a method for seeing if particular zeros ρ satisfy $\text{Re } \rho = \frac{1}{2}$. There are two ideas involved, one of which comes from elementary calculus. If we have a continuous real-valued function $f(x)$ defined on the real numbers, we can sometimes use the intermediate value theorem to count zeros. For example, say $f(1) > 0$, $f(1.7) < 0$, $f(2.3) > 0$. Then we know for sure that f has at least one zero between 1 and 1.7, and at least one zero between 1.7 and 2.3. If we know for other reasons that f has exactly two zeros, then we have accounted for both of them. To locate zeros of the complex function $\zeta(s)$, a real-valued function $g(t)$ is constructed with the property that $\zeta(\frac{1}{2} + it) = 0$ if and only if $g(t) = 0$. By looking at sign changes for $g(t)$ for $0 < t < T$, we can get a *lower bound* for the number of zeros ρ of ζ with $\text{Re } \rho = \frac{1}{2}$ and $0 < \text{Im } \rho < T$. In addition, we can use the so-called *argument principle* from complex analysis to count the *exact number* of zeros with $0 < \text{Im } \rho < T$. If we are lucky and this exact count is equal to our lower bound, then we have accounted for all of ζ 's zeros here, showing that they all have real part $\frac{1}{2}$ (and, in addition, that they are all simple zeros). If the counts did not match, it would not be a disproof of the Riemann hypothesis, but certainly it would indicate a region where we should be checking the data more closely. So far, whenever we have tried this approach, the counts have matched, though sometimes we have been forced to evaluate $g(t)$ at very closely spaced points.

The first few nontrivial zeros were computed by Riemann himself. The famous cryptographer and early computer scientist ALAN TURING [VI.94] also computed some zeta zeros. The current record for this kind of calculation is held by Gourdon, who has shown that the first 10^{13} zeta zeros with positive imaginary part all have real part equal to $\frac{1}{2}$, as predicted by Riemann. Gourdon's method is a modification of that pioneered by Odlyzko and Schönhage (1988), who ushered in the modern age of zeta-zero calculations.

Explicit zeta-function calculations can lead to highly useful explicit prime number estimates. If p_n is the n th prime, then the prime number theorem implies that $p_n \sim n \log n$ as $n \rightarrow \infty$. Actually, there is a secondary term of order $n \log \log n$, and so for all sufficiently large n , we have $p_n > n \log n$. By using explicit zeta estimates, Rosser was able to put a numerical bound on the "sufficiently large" in this statement, and then, by checking small cases, was able to prove that in fact $p_n > n \log n$ for every n . The paper of Rosser and Schoenfeld (1962) is filled with highly useful and numerically explicit inequalities of this kind.

Let us imagine for a moment that the Riemann hypothesis had been proved. Mathematics is never "used up," as there is always that next problem around the bend. Even if we know that all of zeta's nontrivial zeros lie on the line $\text{Im } s = \frac{1}{2}$, we can still ask how they are distributed on this line. We have a fairly concise understanding of how many zeros there should be up to a given height T . In fact, as already found by Riemann, this count is about $(1/2\pi)T \log T$. Thus, on average, the zeros would tend to get closer and closer with about $(1/2\pi) \log T$ of them in a unit interval near height T .

This tells us the average distance, or spacing, between one zeta zero and the next, but there is much more that one can ask about how these spacings are distributed. In order to discuss this question, it is very convenient to "normalize" the spacings, so that the average (normalized) gap between consecutive zeros is 1. By Riemann's result, together with our assumption of the Riemann hypothesis, this can be done if we multiply a gap near T by $(1/2\pi) \log T$, or, equivalently, if for each zero ρ we replace its imaginary part $t = \text{Im } \rho$ by $(1/2\pi)t \log t$. In this way we arrive at a sequence $\delta_1, \delta_2, \dots$ of normalized gaps between consecutive zeros, which on average are about 1.

Checking numerically, we see that some δ_n are large, with others close to 0; it is just the average that is 1. Mathematics is well equipped to study random phenomena, and we have names for various PROBABILITY DISTRIBUTIONS [III.73], such as Poisson, Gaussian, etc. Is this what is happening here? These zeta zeros are not random at all, but perhaps thinking in terms of randomness has promise.

In the early twentieth century, HILBERT [VI.63] and Pólya suggested that the zeros of the zeta function might correspond to the EIGENVALUES [I.3 §4.3] of some OPERATOR [III.52]. Now this is provocative! But what

T&T note: check style later.

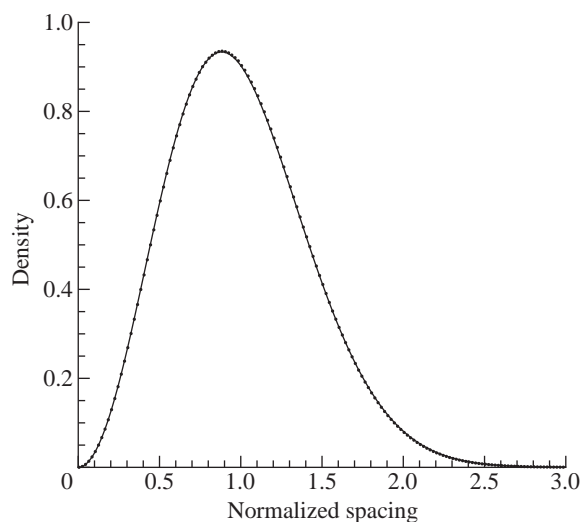


Figure 1 Nearest-neighbor spacing and the Gaudin distribution.

operator? Some fifty years later in a now famous conversation between Dyson and Montgomery at the Institute for Advanced Study, it was conjectured that the nontrivial zeros behave like the eigenvalues of a random matrix from the so-called *Gaussian unitary ensemble*. This conjecture, now known as the GUE conjecture, can be numerically tested in various ways. Odlyzko has done this, and found persuasive evidence for the conjecture: the higher the batches of zeros one looks at, the more closely their distribution corresponds to what the GUE conjecture predicts.

For example, take the 1 041 417 089 numbers δ_n with n starting at $10^{23} + 17\,368\,588\,794$. (The imaginary parts of these zeros are around 1.3×10^{22} .) For each interval $(j/100, (j+1)/100]$ we can compute the proportion of these normalized gaps that lie in this interval, and plot it. If we were dealing with eigenvalues from a random matrix from the GUE, we would expect these statistics to converge to a certain distribution known as the Gaudin distribution (for which there is no closed formula, but which is easily computable). Odlyzko has kindly supplied me with the graph in figure 1, which plots the Gaudin distribution against the data just described (but leaves out every second data point to avoid clutter). Like pearls on a necklace! The fit is absolutely remarkable.

The vital interplay of thought experiments and numerical computation has taken us to what we feel is a deeper understanding of the zeta function. But where

do we go next? The GUE conjecture suggests a connection to random matrix theory, and pursuing further connections seems promising to many. It may be that random matrix theory will allow us only to formulate great conjectures about the zeta function, and will not lead to great theorems. But then again, who can deny the power of a glimpse at the truth? We await the next chapter in this development.

5 Diophantine Equations and the ABC Conjecture

Let us move now from the Riemann hypothesis to FERMAT'S LAST THEOREM [V.12]. Until the last decade it too was one of the most famous unsolved problems in mathematics, once even having a mention on an episode of *Star Trek*. The assertion is that the equation $x^n + y^n = z^n$ has no solutions in positive integers x, y, z, n , where $n \geq 3$. This conjecture had remained unproved for three and a half centuries until Andrew Wiles published a proof in 1995. In addition, perhaps more important than the solution of this particular Diophantine equation (that is, an equation where the unknowns are restricted to the integers), the centuries-long quest for a proof helped establish the field of ALGEBRAIC NUMBER THEORY [IV.1]. And the proof itself established a long-sought and wonderful connection between MODULAR FORMS [III.61] and elliptic curves.

But do you know why Fermat's last theorem is true? That is, just in case you are not an expert on all of the intricacies of the proof, are you surprised that there are in fact no solutions? In fact, there is a fairly simple heuristic argument that supports the assertion. First note that the case $n = 3$, namely $x^3 + y^3 = z^3$, can be handled by elementary methods, and this in fact had already been done by EULER [VI.19]. So, let us focus on the cases when $n \geq 4$.¹ Let S_n be the set of positive n th powers of integers. How likely is it that the sum of two members of S_n is itself a member of S_n ? Well, not at all likely, since Wiles has proved that this never occurs! But recall that we are trying to think naively.

Let us try to mimic our situation by replacing the set S_n with a random set. In fact, we will throw all of the powers together into one set. Following an idea of Erdős and Ulam (1971) we create a set \mathcal{R} by a random process: each integer m is considered independently, and the chance it gets thrown into \mathcal{R} is proportional to $m^{-3/4}$.

1. Actually, Fermat himself had a simple proof in the case $n = 4$, but we ignore this.

This process would typically give us about $x^{1/4}$ numbers in \mathcal{R} in the interval $[1, x]$, or at least this would be the order of magnitude. Now the total number of fourth and higher powers between 1 and x is also about $x^{1/4}$, so we can take our random set \mathcal{R} as modeling the situation for these powers, namely the union of all sets S_n for $n \geq 4$. We ask how likely it is to have $a + b = c$ where a, b , and c all come from \mathcal{R} .

The probability that a number m may be represented as $a + b$ with $0 < a < b < m$ and $a, b \in \mathcal{R}$ is proportional to $\sum_{0 < a < m/2} a^{-3/4} (m - a)^{-3/4}$, since for each a less than m the probability that a and $m - a$ both lie in \mathcal{R} is $a^{-3/4} (m - a)^{-3/4}$. Actually, there is a minor caveat when m is even, since then $a = m - a$ when $a = \frac{1}{2}m$: to cover this, we add the single term $(\frac{1}{2}m)^{-3/4}$ to the above sum. Replacing each $m - a$ in the sum with $\frac{1}{2}m$, we get a larger sum that is easy to estimate and turns out to be proportional to $m^{-1/2}$. That is, the chance that a random number m is a sum of two members of \mathcal{R} is at most a certain quantity that is proportional to $m^{-1/2}$. Now the events that would have to occur for m to be given as such a sum involve numbers smaller than m , so the event that m itself is in \mathcal{R} is independent of these. Therefore, the probability that m is not only the sum of two members of \mathcal{R} , but also itself a member of \mathcal{R} , is at most a quantity proportional to $m^{-1/2} m^{-3/4} = m^{-5/4}$. So now we can count how many times we should expect a sum of two members of \mathcal{R} to itself be a member of \mathcal{R} . This is at most a constant times $\sum_m m^{-5/4}$. But this sum is convergent, so we expect only finitely many examples. Further, since the tail of a convergent series is tiny, we do not expect any large examples.

Thus, this argument suggests that there are at most finitely many positive integer solutions to

$$x^u + y^v = z^w, \quad (2)$$

where the exponents u, v, w are at least 4. Since Fermat's last theorem is the special case when $u = v = w$, we would have at most finitely many counterexamples to that as well.

This seems tidy enough, but now we get a surprise! There are actually *infinitely many solutions* to (2) in positive integers with u, v, w all at least 4. For example, note that $17^4 + 34^4 = 17^5$. This is the case $a = 1, b = 2, u = 4$ of a more general identity: if a, b are positive integers, and $c = a^u + b^u$, we have $(ac)^u + (bc)^u = c^{u+1}$. Another way to get infinitely many examples is to build on the possible existence of just one example. If x, y, z, u, v, w are positive integers satisfying (2), then with

the same exponents, we may replace x, y, z with $a^{vw}x, a^{uw}y, a^{uv}z$ for any integer a , and so get infinitely many solutions.

The point is that events of the kind that we are considering—that a given integer is a power—are not quite independent. For instance, if A and B are both u th powers, then so is AB , and this idea is exploited in the infinite families just mentioned.

So how do we neatly bar these trivialities and come to the rescue of our heuristic argument? One simple way to do this is to insist that the numbers x, y, z in (2) be relatively prime. This gives no restriction whatsoever in the Fermat case of equal exponents, since a solution to $x^n + y^n = z^n$ with d the greatest common divisor of x, y, z leads to the coprime solution $(x/d)^n + (y/d)^n = (z/d)^n$.

Concerning Fermat's last theorem, one might ask how far it had actually been verified before the final proof by Wiles. The paper by Buhler et al. (1993) reports a verification for all exponents n up to 4 000 000. This type of calculation, which is far from trivial, has its roots in nineteenth-century work of KUMMER [VI.40] and early-twentieth-century work of Vandiver. In fact, Buhler et al. (1993) also verify in the same range a related conjecture of Vandiver dealing with cyclotomic fields, but this conjecture may in fact be false in general.

The probabilistic thinking above, combined with computation of small cases, can carry us deeply into some very provocative conjectures. The above probabilistic argument can easily be extended to suggest that (2) has at most finitely many relatively prime solutions x, y, z over all possible exponent triples u, v, w with $1/u + 1/v + 1/w < 1$. This conjecture has come to be known as the Fermat–Catalan conjecture, since it contains within it essentially Fermat's last theorem and also the Catalan conjecture (recently proved by Mihăilescu) that 8 and 9 are the only consecutive powers.

It is good that we do allow for the possibility that there are *some* solutions, and this is where our main topic of computing comes in. For example, since $1 + 8 = 9$, we have a solution to $x^7 + y^3 = z^2$, where $x = 1, y = 2$, and $z = 3$. (The exponent 7 is chosen to insure that the reciprocal sum of the exponents is less than 1. Of course, we could replace 7 by any larger integer, but since in each case the power involved is the number 1, they should all together be considered as just one

example.) Here are the known solutions to (2):

$$\begin{aligned}
1^n + 2^3 &= 3^2, \\
2^5 + 7^2 &= 3^4, \\
13^2 + 7^3 &= 2^9, \\
2^7 + 17^3 &= 71^2, \\
3^5 + 11^4 &= 122^2, \\
33^8 + 1\,549\,034^2 &= 15\,613^3, \\
1414^3 + 2\,213\,459^2 &= 65^7, \\
9262^3 + 15\,312\,283^2 &= 113^7, \\
17^7 + 76\,271^3 &= 21\,063\,928^2, \\
43^8 + 96\,222^3 &= 30\,042\,907^2.
\end{aligned}$$

The larger members were found in an exhaustive computer search by Beukers and Zagier. Perhaps this is the complete list of all solutions, or perhaps not—we have no proof.

However, for particular choices of u, v, w , more can be said. Using results from a famous paper of Faltings, Darmon and Granville (1995) have shown that for any fixed choice of u, v, w with reciprocal sum at most 1, there are at most finitely many coprime triples x, y, z solving (2). For a particular choice of exponents, one might try to actually find all of the solutions. If it can be handled at all, this task can involve a delicate interplay between ARITHMETIC GEOMETRY [IV.5], effective methods in transcendental number theory, and good hard computing. In particular, the exponent triple sets $\{2, 3, 7\}$, $\{2, 3, 8\}$, $\{2, 3, 9\}$, and $\{2, 4, 5\}$ are known to have all their solutions in the above table. See Poonen et al. (2007) for the treatment of the case $\{2, 3, 7\}$ and links to other work.

THE ABC CONJECTURE [V.1] of Oesterlé and Masser is deceptively simple. It involves positive integer solutions to the equation $a + b = c$, hence the name. To put some meaning into $a + b = c$, we define the *radical* of a nonzero integer n as the product of the primes that divide n , denoting this as $\text{rad}(n)$. So, for example, $\text{rad}(10) = 10$, $\text{rad}(72) = 6$, and $\text{rad}(65\,536) = 2$. In particular, high powers have small radicals in comparison to the number itself, and so do many other numbers. Basically, the ABC conjecture asserts that if $a + b = c$, then the radical of abc cannot be too small. More specifically we have the following.

The ABC conjecture. For each $\varepsilon > 0$ there are at most finitely many relatively prime positive integer triples a, b, c with $a + b = c$ and $\text{rad}(abc) < c^{1-\varepsilon}$.

Note that the ABC conjecture immediately solves the Fermat–Catalan problem. Indeed, if u, v, w are positive integers with $1/u + 1/v + 1/w < 1$, then it is easily found that we must have $1/u + 1/v + 1/w \leq 41/42$. Suppose we have a coprime solution to (2). Then $x \leq z^{w/u}$ and $y \leq z^{w/v}$, so that

$$\text{rad}(x^u y^v z^w) \leq xyz \leq (z^w)^{41/42}.$$

Thus, the ABC conjecture with $\varepsilon = 1/42$ implies that there are at most finitely many solutions.

The ABC conjecture has many other marvelous consequences; for a delightful survey, see Granville and Tucker (2002). In fact, the ABC conjecture and its generalizations can be used to prove so many things that I have joked that it is beginning to resemble a false statement, since a false statement implies everything. But probably the ABC conjecture is true. Indeed, though a bit harder to see, the Erdős–Ulam probabilistic argument can be modified to provide heuristic evidence for it too.

Basic to this argument is a perfectly rigorous result on the distribution of integers n for which $\text{rad}(n)$ is below some bound. These ideas, which lead to a more explicit version of the ABC conjecture, are worked through in the thesis of van Frankenhuijsen and by Stewart and Tenenbaum. Here is a slightly weaker statement: if $a + b = c$ are relatively prime positive integers and c is sufficiently large, then we have

$$\text{rad}(abc) > c^{1-1/\sqrt{\log c}}. \quad (3)$$

One might wonder how the numerical evidence stacks up against (3). This inequality asserts that if $\text{rad}(abc) = r$, then $\log(c/r)/\sqrt{\log c} < 1$. So, let $T(a, b, c)$ denote the test statistic $\log(c/r)/\sqrt{\log c}$. A Web site maintained by Nitaj (www.math.unicaen.fr/~nitaj/abc.html) contains a wealth of information about the ABC conjecture. Checking the data, there are quite a few examples with $T(a, b, c) \geq 1$, the champion so far being

$$\begin{aligned}
a &= 7^2 \cdot 41^2 \cdot 311^3 = 2\,477\,678\,547\,239 \\
b &= 11^{16} \cdot 13^2 \cdot 79 = 613\,474\,843\,408\,551\,921\,511 \\
c &= 2 \cdot 3^3 \cdot 5^{23} \cdot 953 = 613\,474\,845\,886\,230\,468\,750 \\
r &= 2 \cdot 3 \cdot 5 \cdot 7 \cdot 11 \cdot 13 \cdot 41 \cdot 79 \cdot 311 \cdot 953 \\
&= 28\,828\,335\,646\,110,
\end{aligned}$$

so that

$$T(a, b, c) = \frac{\log(c/r)}{\sqrt{\log c}} = 2.43886\dots$$

Is it always true that $T(a, b, c) < 2.5$?

One can get carried away with heuristics, forgetting that one is not actually proving a theorem, but making a guess. Heuristics are often based on the idea of randomness, and all bets are off if there is some underlying structure. But how do we know that there is no underlying structure? Consider the case of an “ $abcd$ conjecture.” Here we consider integers a , b , c , and d with $a + b + c + d = 0$. The condition that the terms be relatively prime now takes on two possible meanings: pairwise relatively prime or no nontrivial common divisor of all four numbers. The first condition seems more in the spirit of the three-term conjecture, but may be a tad too strong in that it disallows using any even numbers. So say we take the four terms with no pair having a common factor greater than 2. Under this condition, our heuristics seem to suggest that for each $\varepsilon > 0$, we have

$$\text{rad}(abcd)^{1+\varepsilon} < \max\{|a|, |b|, |c|, |d|\} \quad (4)$$

for at most finitely many cases. But consider the polynomial identity

$$(x+1)^5 = (x-1)^5 + 10(x^2+1)^2 - 8$$

(suggested to me by Granville). If we take x as a multiple of 10, the four terms involved in the identity are pairwise relatively prime except for the last two, which have a common factor of 2. Let $x = 11^k - 1$, which is a multiple of 10. The largest of the four terms is 11^{5k} , and the radical of the product of the four terms is at most

$$110(11^k - 2)((11^k - 1)^2 + 1) < 110 \cdot 11^{3k}.$$

The heuristics are saying that this cannot be, yet here it is right before our eyes!

What is happening is that the polynomial identity is supplying an underlying structure. For the four-term $abcd$ conjecture, Granville conjectures that for each $\varepsilon > 0$, all counterexamples to (4) come from at most finitely many polynomial families. And the number of polynomial families grows to infinity as ε shrinks to 0.

We have looked here at only a small portion of the field of Diophantine equations, and then we have looked mainly at the dynamic relationship between heuristics and computational searches for small solutions. For much more on the subject of computational Diophantine methods, see Smart (1998).

Heuristic arguments often assume that the objects of study behave as if they were random, and we have visited several cases where it is useful to think this way. Other examples include the twin-prime conjecture (there are infinitely many primes p such that $p + 2$

is prime), Goldbach’s conjecture (every even number larger than 2 is the sum of two primes), and countless other conjectures in number theory. Often the computational evidence for the probabilistic view is striking, even overwhelming, and we become convinced of the truth of our model. But on the other hand, if it is this pseudo-proof that is all we have to go on, we may still be very far from the truth. Nevertheless, the interplay of computations and heuristic thinking forms an indispensable part of our arsenal, and mathematics is the richer for it.

Remarks and Acknowledgments

I would like to recommend to the reader the book by Cohen (1993) for a discussion of computational algebraic number theory, a subject that is neglected in this article. I am grateful to the following people, who generously shared their expertise: X. Gourdon, A. Granville, A. Odlyzko, E. Schaefer, K. Soundararajan, C. Stewart, R. Tijdeman, and M. van Frankenhuijsen. I am also thankful to A. Granville and D. Pomerance for helpful suggestions with the exposition. I was supported in part by NSF grant DMS-0401422.

Further Reading

- Agrawal, M., N. Kayal, and N. Saxena. 2004. PRIMES is in P. *Annals of Mathematics* 160:781–93.
- Buhler, J., R. Crandall, R. Ernvall, and T. Metsänkylä. 1993. Irregular primes and cyclotomic invariants to four million. *Mathematics of Computation* 61:151–53.
- Cohen, H. 1993. *A Course in Computational Algebraic Number Theory*. Graduate Texts in Mathematics, volume 138. New York: Springer.
- Crandall, R., and C. Pomerance. 2005. *Prime Numbers: A Computational Perspective*, 2nd edn. New York: Springer.
- Darmon, H., and A. Granville. 1995. On the equations $z^m = F(x, y)$ and $Ax^p + By^q = Cz^r$. *Bulletin of the London Mathematical Society* 27:513–43.
- Erdős, P., and S. Ulam. 1971. Some probabilistic remarks on Fermat’s last theorem. *Rocky Mountain Journal of Mathematics* 1:613–16.
- Granville, A., and T. J. Tucker. 2002. It’s as easy as abc . *Notices of the American Mathematical Society* 49:1224–31.
- Odlyzko, A. M., and A. Schönhage. 1988. Fast algorithms for multiple evaluations of the Riemann zeta function. *Transactions of the American Mathematical Society* 309: 797–809.
- Poonen, B., E. Schaefer, and M. Stoll. 2007. Twists of $X(7)$ and primitive solutions to $x^2 + y^3 = z^7$. *Duke Mathematics Journal* 137:103–58.

Rosser, J. B., and L. Schoenfeld. 1962. Approximate formulas for some functions of prime numbers. *Illinois Journal of Mathematics* 6:64–94.

Smart, N. 1998. *The Algorithmic Resolution of Diophantine Equations*. London Mathematical Society Student Texts, volume 41. Cambridge: Cambridge University Press.

IV.4 Algebraic Geometry

János Kollár

1 Introduction

Succinctly put, algebraic geometry is the study of geometry using polynomials and the investigation of polynomials using geometry.

Many of us were taught the beginnings of algebraic geometry in high school, under the name “analytic geometry.” When we say that $y = mx + b$ is the equation of a line L , or that $x^2 + y^2 = r^2$ describes a circle C of radius r , we establish a basic connection between geometry and algebra.

If we want to find the points where the line L and the circle C intersect, we just substitute $mx + b$ for y in the circle equation to get $x^2 + (mx + b)^2 = r^2$ and solve the resulting quadratic equation to obtain the x coordinates of the two intersection points.

This simple example encapsulates the method of algebraic geometry: a geometric problem is translated into algebra, where it is readily solvable; conversely, we get insight into algebra problems by using geometry. It is hard to guess the solutions of systems of polynomial equations, but once a corresponding geometric picture is drawn, we start to have a qualitative understanding of them. The precise quantitative answer is then provided by algebra.

2 Polynomials and Their Geometry

Polynomials are the expressions one can put together from variables and numbers by addition and multiplication. The most familiar are one-variable polynomials such as $x^3 - x + 4$, but we can use two or three variables to get, for instance, $2x^5 - 3xy^2 + y^3$ (which has degree 5 in two variables) or $x^5 - y^7 + x^2z^8 - xyz + 1$ (which has degree 10 in three variables). In general, one can use n variables, in which case they are frequently denoted by x_1, x_2, \dots, x_n , and we write $f(x_1, \dots, x_n)$, $f(\mathbf{x})$ or simply f to denote an unspecified polynomial.

Polynomials are the only functions that computers can work with. (Although your pocket calculator is

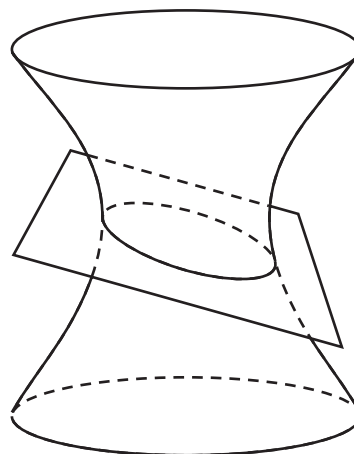


Figure 1 A hyperboloid intersecting a plane.

likely to have a button for logarithms, it is secretly computing a polynomial whose value at a number b agrees with $\log b$ up to many decimal places.)

We can slightly rewrite the equations we gave earlier for the line L and the circle C : as $y - mx - b = 0$ and $x^2 + y^2 - r^2 = 0$. We can then describe L and C as *zero sets*: L is the zero set of $y - mx - b$ (that is, the set of all points (x, y) such that $y - mx - b = 0$) and C is the zero set of $x^2 + y^2 - r^2$.

Similarly, the zero set of $2x^2 + 3y^2 - z^2 - 7$ in 3-space is a hyperboloid, the zero set of $z - x - y$ in 3-space is a plane, and the common zero set of these two equations in 3-space is the intersection of the hyperboloid and the plane, which is an ellipse (see figure 1).

The set of common zeros of a system of polynomial equations in any number of variables is called an *algebraic set*. These are the basic objects of algebraic geometry.

Most people feel that geometry ends in 3-space. Very few have a feeling for 4-space, also called *space-time*, and 5-space is by and large inconceivable to almost everyone. So what is the meaning of geometry in many variables?

Algebra comes to our rescue here. While I have great difficulty visualizing what a four-dimensional sphere of radius r in 5-space should be, I can easily write down its equation,

$$x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 - r^2 = 0,$$

and work with it. This equation is also something a computer can handle, which is immensely useful in applications.

I will, nonetheless, stick to two or three variables for the rest of this article. This is where all geometry starts and there are plenty of interesting questions and results.

The importance of algebraic geometry derives from the fact that significant interactions between algebra and geometry happen very frequently. Let us look at two examples, just for illustration.

3 Most Shapes Are Algebraic

Shapes that occur frequently enough to have their own name, for instance, lines, planes, circles, ellipses, hyperbolas, parabolas, hyperboloids, paraboloids, ellipsoids, are almost all algebraic. Even the more esoteric conchoid (or shell curve) of Dürer, the trident of NEWTON [VI.14], and the folium of Kepler are algebraic.

Some shapes cannot be described by polynomial equations, but they can be described by polynomial inequalities. For instance, the inequalities $0 \leq x \leq a$ and $0 \leq y \leq b$ together describe a rectangle with side lengths a, b . Shapes described by polynomial inequalities are called *semi-algebraic*, and every polyhedron is semi-algebraic.

Not everything is an algebraic set, though. Look, for example, at the graph of the sine function $y = \sin x$. This crosses the x -axis infinitely many times (at multiples of π). If $f(x)$ is any polynomial, then it has at most as many roots as its degree, so $y = f(x)$ will never look like $y = \sin x$.

We can, however, get very close to $\sin x$ with a polynomial if we concentrate on values of x that are not too large. For instance, the degree-7 Taylor polynomial

$$x - \frac{1}{6}x^3 + \frac{1}{120}x^5 - \frac{1}{5040}x^7$$

differs from $\sin x$ by an error of at most 0.1 for $-\pi < x < \pi$. This is a very special case of a basic theorem of Nash that says that every “reasonable” geometric shape is algebraic if we ignore what happens very far from the origin. So, what is reasonable? Certainly not everything. Fractals seem profoundly nonalgebraic. The nicest shapes are MANIFOLDS [I.3 §6.9], and all of these can be described by polynomials.

Nash’s theorem. *Let M be any manifold in \mathbb{R}^n . Fix any large number R . Then there is a polynomial f whose zero set is as close to M as we want, at least inside a ball of radius R around the origin.*

4 Codes and Finite Geometries

Consider the equation $x^2 + y^2 = z^2$, which describes a double cone in 3-space (see figure 4). If we confine ourselves to natural numbers, then the solutions of $x^2 + y^2 = z^2$ are the *Pythagorean triples*, corresponding to right-angled triangles where all sides have integer lengths, of which the two best-known examples are (3, 4, 5) and (5, 12, 13).

Let us now look at the same equation, but declare that we care only about the *parities* of the two sides (that is, whether they are even or odd). For instance, $3^2 + 15^2$ and 4^2 are both even, so we say that $3^2 + 15^2 \equiv 4^2 \pmod{2}$ (see MODULAR ARITHMETIC [III.60]). The parities of $x^2 + y^2$ and of z^2 depend only on those of x, y , and z , so we can pretend that x, y , and z are all either 0 (the even case) or 1 (the odd case). Our equation modulo 2 therefore has four solutions:

$$000, 011, 101, 110.$$

These look like code words in a computer message. It was quite a surprise when it was discovered that using polynomials and their solutions modulo 2 is a great—probably the best—way of constructing error-correcting codes (see RELIABLE TRANSMISSION OF INFORMATION [VII.6]).

There is something very substantial and new happening here. Let us think for a moment about what 3-space is for us. For many it is an amorphous everything, but for algebraic geometers (with DESCARTES [VI.11] as our ancestor) it is simply a collection of points described by three numbers, the x, y , and z coordinates. Let us make a jump here, and declare that “3-space modulo 2” is the collection of all “points” given by three coordinates modulo 2. Four of these are listed above, and there are four more. The beauty of algebra is that suddenly we can talk about lines, planes, spheres, cones in this “3-space having only eight points.”

We do not need to stop here, and one can work modulo any integer. For example, working modulo 7, we have 0, 1, 2, 3, 4, 5, 6 as possible coordinates, and so “3-space modulo 7” has $7^3 = 343$ points.

Talking about geometry in these spaces is very intriguing, but also technically difficult. Its great reward is that one can view this process as a “discretization” of ordinary space. Working modulo n for large n (especially when n is a prime number) gets very close to the usual geometry.

This approach is especially fruitful in number-theoretic questions. It was, for instance, instrumental in Wiles’s proof of Fermat’s last theorem.

For more on these topics, see ARITHMETIC GEOMETRY [IV.5].

5 Snapshots of Polynomials

Consider the equation $x^2 + y^2 = R$. If $R > 0$, then the real solutions form a circle of radius \sqrt{R} ; if $R = 0$, we get only the origin; and if $R < 0$, we get the empty set. Thus, if $R > 0$, then the geometry of the solution set determines what R is, but otherwise it does not. We can of course look at complex solutions, and the complex solutions always determine R . (For instance, the intersection points with the x -axis are $(\pm\sqrt{R}, 0)$.)

If R is a rational number, we can ask about rational solutions of $x^2 + y^2 = R$, and if R is an integer, we can also look for solutions in the “plane modulo m ” for any m .

One can even look for solutions where $x = x(t)$, $y = y(t)$ are themselves polynomials in a variable t . (Most generally, we can ask for solutions where x, y are elements of any ring containing the number R .)

To my mind, the polynomial is the central object, and each time we look at solution sets we are taking a “snapshot” of the polynomial. Some snapshots are good (like the above real snapshot for $R > 0$) and some are bad (like the above real snapshot for $R < 0$).

How good can snapshots be? Can we determine a polynomial from its snapshots?

One frequently talks about “the” equation of a hyperbola, but “an” equation would be more correct. Indeed, the hyperbola $x^2 - y^2 - R = 0$ can also be given by an equation $cx^2 - cy^2 - cR = 0$, for any $c \neq 0$. We can also use the equation $(x^2 - y^2 - R)^2 = 0$, which we may well not recognize in its expanded form. Higher powers can also be used. What about the equation $f(x, y) = (x^2 - y^2 - R)(x^2 + y^2 + R^2) = 0$? If we look only at real solutions, this is still just the hyperbola since $x^2 + y^2 + R^2$ is always positive for x, y real. However, as with one-variable polynomials, one should look at all complex roots to understand everything. Then we see that $f(\sqrt{-1}R, 0) = 0$, but the complex point $(\sqrt{-1}R, 0)$ is not on the hyperbola $x^2 - y^2 - R = 0$. In general, as long as $R \neq 0$, we get that if $f(x, y)$ is a polynomial that has exactly the same complex roots as $x^2 - y^2 - R$, then $f = c(x^2 - y^2 - R)^m$ for some m and $c \neq 0$.

Why is the $R = 0$ case different? The reason is that for $R \neq 0$ the polynomial $x^2 - y^2 - R$ is *irreducible* (that is, it cannot be written as the product of other polynomials), while $x^2 - y^2 = (x + y)(x - y)$ is reducible

with *irreducible factors* $x + y$ and $x - y$. In the latter case one gets that if $g(x, y)$ is a polynomial that has exactly the same complex roots as $x^2 - y^2$, then $f = c \cdot (x + y)^m(x - y)^n$ for some m, n and $c \neq 0$.

The analogous question for systems of equations is answered by the fundamental theorem of algebraic geometry. It is sometimes called Hilbert’s theorem on the zeros, but its German name is used most of the time. For simplicity, we state only the case of one equation.

Hilbert’s Nullstellensatz. *Two complex polynomials f and g have the same complex solutions if and only if they have the same irreducible factors.*

We can do even better for polynomials with integer coefficients. For instance, $x^2 - y^2 - 1 = 0$ and $2(x^2 - y^2 - 1) = 0$ have the same solutions over the real or complex numbers, and the same solutions modulo p for any odd prime p , but they have different solutions modulo 2. The general result in this case is easy and simple.

Arithmetic Nullstellensatz. *Two polynomials with integer coefficients f and g have the same solutions modulo m for every m if and only if $f = \pm g$.*

6 Bézout’s Theorem and Intersection Theory

If $h(x)$ is a polynomial of degree n , then it has n complex roots, at least when they are counted with multiplicity. What happens with a system $f(x, y) = g(x, y) = 0$? Geometrically we see two curves in the plane, so we expect that there will typically be finitely many intersection points.

If f, g are both linear, we have two lines in the plane. These usually intersect in a single point, but they can be parallel and they can coincide. The first case leads to the classical declaration that “parallel lines meet at infinity” and the definition of projective planes and PROJECTIVE SPACES [III.74]. (The introduction of projective spaces and the corresponding projective varieties is a key step in algebraic geometry. It is somewhat technical so we shall skip it here, but it is indispensable even at the most basic level.)

Next, consider two polynomials of degree 2, that is, two plane conics. Two smooth conics usually intersect in at most four points (just try this by drawing two ellipses). There are also some rather degenerate cases. Two conics may coincide, or, if they are both reducible, they can have a common line. In any case, we are ready to formulate a basic result, dating back to 1779.

Bézout's theorem. Let $f_1(\mathbf{x}), \dots, f_n(\mathbf{x})$ be n polynomials in n variables, and for each i let d_i be the degree of f_i . Then either

- (i) the equation(s) $f_1(\mathbf{x}) = \dots = f_n(\mathbf{x}) = 0$ have at most $d_1 d_2 \dots d_n$ solutions; or
- (ii) the f_i vanish identically on an algebraic curve C , and so there is a continuous family of solutions.

As an example, the second alternative happens for the system of equations $xz - y^2 = y^3 - z^2 = x^3 - z = 0$, which has (t, t^2, t^3) as a solution for any t . This case is actually quite rare. If we pick the coefficients of the polynomials f_i randomly, then the first alternative happens with probability 1.

Ideally, we would like to make the stronger claim that if the first alternative happens, then there are *exactly* $d_1 d_2 \dots d_n$ solutions, but counted “with multiplicity.” This actually works, and gives us our first example of an extremely useful feature of algebraic geometry. Even in very degenerate situations it is possible to define and count the multiplicities easily. This is frequently of great help since the typical (or “generic”) cases are usually very hard to compute. To get around this problem, we can sometimes find a special, degenerate case where we know that the answer will be the same, but the computations are much easier.

There are two ways to think about multiplicity: one algebraic and one geometric. The algebraic definition is computationally very efficient, but somewhat technical. The geometric interpretation is easier to explain, so that is the one we shall give here, but it would be hard to compute with in practice.

If $\mathbf{x} = \mathbf{p}$ is an isolated solution of the equations $f_1(\mathbf{x}) = \dots = f_n(\mathbf{x}) = 0$ with multiplicity m , then the perturbed system

$$f_1(\mathbf{x}) + \epsilon_1 = \dots = f_n(\mathbf{x}) + \epsilon_n = 0$$

has exactly m solutions near $\mathbf{x} = \mathbf{p}$ for almost all small values of the ϵ_i .

Intersection theory is the branch of algebraic geometry that deals with generalizations of Bézout's theorem. Above, we looked at intersections of *hypersurfaces*—that is, of zero sets of single polynomials—but we may wish to look at intersections of more general algebraic sets. Also, even when the second alternative holds, we may want to count the number of isolated intersection points; this can be very tricky but also very useful.

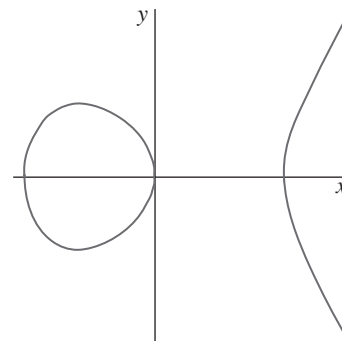


Figure 2 A smooth cubic: $y^2 = x^3 - x$.

7 Varieties, Schemes, Orbifolds, and Stacks

Consider the system $xz = yz = 0$ in 3-space. It consists of two pieces, the $z = 0$ plane and the $x = y = 0$ line. It is easy to see that neither the plane nor the line can be written as the union of algebraic sets (except by nitpickers who point out that the line is the union of the line itself and of any point on the line). In general, any algebraic set can be written in exactly one way as the union of smaller algebraic sets that in turn cannot be decomposed further. These basic building blocks are called *irreducible algebraic sets* or *algebraic varieties*.

Sometimes this is not exactly what one would naively expect. For instance, the curve in figure 2 has two connected components. The two parts are, however, not algebraic sets.

An explanation is provided by looking at the *complex* solutions of this equation. We shall see later that these form a connected set, namely a torus (with a missing point at infinity). We see two components when we look at the real solutions because we are taking a cross-section of this torus.

In general, the zero set $f = 0$ is irreducible as an algebraic set if and only if f is irreducible as a polynomial (or if it is the power of an irreducible polynomial). The implication in one direction is easy to see: if $f = gh$, then the zero set of f is the union of the zero set of g and of the zero set of h .

For many questions, keeping track only of the zero set is not enough. For instance, look at the polynomial $f = x^2(x - 1)(x - 2)^3$. It has degree 6 and three roots at $x = 0, 1, 2$. These roots behave differently, however, and one usually says that f has a double root at $x = 0$ and a triple root at $x = 2$. If we perturb f by adding a small number ϵ to it, then the perturbed equation $f(x) + \epsilon = 0$ has two (complex) solutions near 0, one

solution near 1 and three (complex) solutions near 2. Thus, these multiplicities carry important geometric meaning about the perturbation of the equation.

Similarly, it is natural to say that while $x^2y = 0$ and $xy^3 = 0$ define the same algebraic set (consisting of the two axes), the first “assigns multiplicity 2” to the y -axis and the other “assigns multiplicity 3” to the x -axis.

More complicated things can happen for systems of equations. Consider the systems $x = y^2 = 0$ and $x^3 = y = 0$ in 3-space. Both define the z -axis and it is reasonable to say that the first does so with multiplicity 2, the second with multiplicity 3. There is, however, a further difference. In the first case the multiplicity seems to “go in the y -direction” and in the second case it seems to go in the x -direction. We can also look at other systems, like $x - cy = y^3 = 0$, if we want to see more complicated behavior.

Roughly speaking, a *scheme* is an algebraic set where we also keep track of the multiplicities and of the directions they occur in.

Consider the xy -plane and consider the map that reflects across the origin. Thus a point (x, y) is mapped to $(-x, -y)$. Let us try to glue each point (x, y) to its image $(-x, -y)$. What do we get? The right half-plane $x \geq 0$ is mapped to the left half-plane $x \leq 0$, so it is enough to work out what happens with the right half-plane. The positive y -axis is glued to the negative y -axis, and the resulting surface is a dunce cap (but less pointy).

Algebraically, it is one half of the cone $z^2 = x^2 + y^2$. This cone looks nice and smooth except at the vertex. There it is more complicated, but the above construction shows that it can be obtained from a plane by a reflection across a point. More generally, suppose we take the n -dimensional space \mathbb{R}^n and finitely many symmetries of it. If we glue together points that move into each other, we again get an algebraic variety, most of whose points are smooth, but some of which are more complicated. A variety made up of pieces like these is called an *orbifold*. (When this is defined more precisely, we also keep track of which symmetries have been used.) In practice, such varieties occur frequently; that is why they deserve a separate name.

Finally, if we marry a scheme to an orbifold, the outcome is a *stack*. The study of stacks is strongly recommended to people who would have been flagellants in earlier times.

8 Curves, Surfaces, Threefolds

As with any geometric object, one of the simplest questions one can ask about a variety is: what is its dimension? As expected, a curve in the plane has dimension 1, and a surface in 3-space has dimension 2. This seems quite simple until one writes down examples like $S = (x^4 + y^4 + z^4 = 0)$, which is only the origin in \mathbb{R}^3 . This example is, nonetheless, still two dimensional: the explanation is that we were looking at the wrong snapshot. Using complex numbers we can solve the equation as $z = \sqrt[4]{-x^4 - y^4}$, so the complex solutions of $x^4 + y^4 + z^4 = 0$ can be described by two independent variables x, y and a dependent variable z . Thus, it is quite reasonable to say that S is two dimensional.

This idea works more generally. If X is any variety in some complex space \mathbb{C}^n , then choose a random set of n independent directions to serve as a basis, or coordinate system, for \mathbb{C}^n , and hence for X . With probability 1 (i.e., except in degenerate cases) one finds that there is some d such that the first d coordinates of a point x in X can vary independently, while the rest depend on them. This number d depends on X only and is called the *dimension* (or, to be precise, the *algebraic dimension*) of X .

If X is a variety and f is a polynomial, then the intersection $X \cap (f = 0)$ has dimension one less than $\dim X$ (unless f vanishes identically on X or never takes the value zero on X).

If X is a subset of \mathbb{R}^n defined by real equations, and if it is smooth (see the next section for a discussion of smoothness), then its topological dimension (see DIMENSION [III.17]) is the same as its algebraic dimension.

For complex varieties, the topological dimension is twice the algebraic dimension. Thus, for an algebraic geometer, \mathbb{C}^n has dimension n . In particular, for us \mathbb{C} is the “complex line,” whereas everybody else calls this the “complex plane.” Our “complex plane” is, of course, \mathbb{C}^2 .

A variety of dimension 1 is called a *curve*. A *surface* is a variety of dimension 2, and a *threefold* is a variety of dimension 3.

The theory of algebraic curves is a very well developed and beautiful subject. We shall see later how one can start to get an overview of all algebraic curves. Surfaces have been intensively studied for the last century, and now we have reached a reasonably complete understanding of them. This is a much more complicated theory than for curves. Still very little is known for

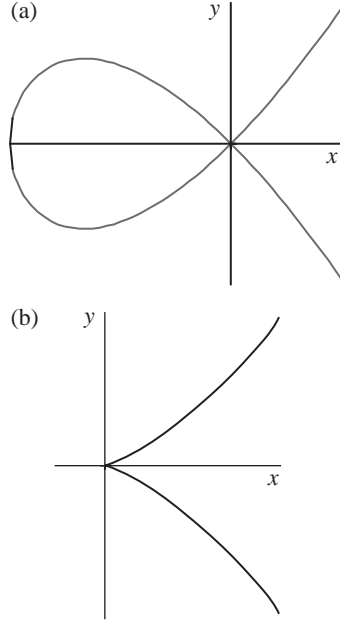


Figure 3 Singular cubics: (a) $y^2 = x^3 + x^2$ and (b) $y^2 = x^3$.

varieties of dimension 3 and up. At least conjecturally, all these dimensions behave in roughly the same way. Despite some progress, especially in dimension 3, many questions are wide open.

9 Singularities and Their Resolutions

If we look at the simplest examples of algebraic curves in figure 3, we see that most points of a curve are smooth, but that there may be a finite set of more complicated singular points. Let us compare these with the curve in figure 2.

All three curves pass through the origin, since their equation has no constant term. The equation of figure 2 has a linear term and the curve looks nice and smooth at the origin, whereas the equations of figure 3 contain no linear term and the curves are more complicated at the origin. This is not an accident. For small values of x , the higher powers x^2, x^3, \dots are much smaller than x in absolute value, so near the origin the linear terms dominate. If we have only linear terms $ax + by = 0$, we get a line through the origin, and an algebraic curve $ax + by + cx^2 + gxy + ey^2 + \dots = 0$ is close to the line $ax + by = 0$, at least for very small values of x and y .

The study of a curve near another point with coordinates (p, q) can be reduced to the case $(p, q) = (0, 0)$ via the coordinate change $(x, y) \mapsto (x - p, y - q)$.

In general, if $f(\mathbf{0}) = 0$ and f has a (nonzero) linear term $L(f)$, the hypersurface $f = 0$ is very close to the hyperplane $L(f) = 0$. This is the so-called *implicit function theorem*. Such points are called *smooth*. Points that are not smooth are called *singular*. One can easily show that the singular points of X form an algebraic set, defined by the vanishing of all partial derivatives $\partial f / \partial x_i$. A random hypersurface will, with probability 1, be smooth, but there are many singular hypersurfaces as well.

The smooth and singular points of an arbitrary variety of dimension d can be defined analogously by comparing X with d -dimensional linear subspaces.

Singularities also occur in other geometric fields, such as topology and differential geometry, but by and large these fields shy away from their study (with the notable exception of catastrophe theory). By contrast, algebraic geometry provides very powerful tools for their investigation.

Let us start with singularities of hypersurfaces, or equivalently with *critical points* of functions. When thinking about these it is natural to work not just with polynomials but with more general power series, that is, functions $f(x_1, \dots, x_n)$ that can be written as “polynomials of infinite degree.” For simplicity of notation we shall assume that $f(\mathbf{0}) = 0$. Two functions f, g are considered to be *equivalent* if there is a coordinate change $x_i \mapsto \phi_i(\mathbf{x})$, where each ϕ_i is given by a power series, such that $f(\phi_1(\mathbf{x}), \dots, \phi_n(\mathbf{x})) = g(\mathbf{x})$.

In the one-variable case, any f can be written as

$$f = x^m(a_m + a_{m+1}x + \dots),$$

where $a_m \neq 0$. The (inverse of the) substitution

$$x \mapsto x \sqrt[m]{a_m + a_{m+1}x + \dots}$$

then shows that f is equivalent to x^m . The functions x^m are inequivalent for different values of m , so in this particular case the lowest-degree monomial occurring in f determines f up to equivalence. (Note that even if f is a polynomial, the above change of variable involves an infinite power series: it is because we cannot invert polynomials, even locally, that it is more convenient to consider general power series.)

In general, the lowest-degree terms of a power series do not determine the singularity, but taking more terms is usually enough to do so, because of the following result.

Algebraization of analytic singularities. *Given a power series f , let $f_{\leq N}$ denote the polynomial obtained from f by deleting all monomials of degree greater than N .*

If $\mathbf{0}$ is an isolated singular point of the hypersurface ($f = 0$), then f is equivalent to $f_{\leq N}$ for sufficiently large N .

To see an example of a nonisolated singularity at $\mathbf{0}$, take

$$\begin{aligned} g(x, y, z) &= \left(y + \frac{x}{1-x}\right)^2 - z^3 \\ &= (y + x + x^2 + x^3 + \cdots)^2 - z^3. \end{aligned}$$

It has singular points not just at $\mathbf{0}$, but everywhere along the curve $y + (x/(1-x)) = z = 0$. On the other hand, one can easily check that all truncations $g_{\leq N}$ do have an isolated singular point at $\mathbf{0}$.

If we have two power series, f and g , we can view functions of the form $f + \epsilon g$ as perturbations of f . A very fruitful question of singularity theory asks: what can we say about the perturbations of a given polynomial or power series f ?

For instance, in the one-variable case, the polynomial x^m can be perturbed as $x^m + \epsilon x^r$, which is equivalent to x^r if $r < m$. Every perturbation contains x^m , so if $r > m$, then no perturbation of x^m will be equivalent to x^r (because near the origin x^m will be much larger than x^r). Hence, up to equivalence, the set of all possible perturbations of x^m is $\{x^r : r \leq m\}$.

On the other hand, it is not hard to see that for any given ϵ , there are only twenty-four different values of η for which the polynomials $x y (x^2 - y^2) + \epsilon y^2 (x^2 - y^2)$ and $x y (x^2 - y^2) + \eta y^2 (x^2 - y^2)$ are equivalent. (Indeed, both polynomials describe four lines through the origin. The first one gives the lines $y = 0$, $x = y$, $x = -y$, and $x = -\epsilon y$, and the second gives the same lines except that η replaces ϵ . The linear part of any supposed equivalence gives a linear transformation mapping the first set of four lines to the second. There are twenty-four ways to assign which line goes to which line.) Thus $x y (x^2 - y^2)$ has a continuous family of inequivalent perturbations.

Simple singularities. Suppose that the polynomial or power series $f(x_1, \dots, x_n)$ has only finitely many inequivalent perturbations. Then f is equivalent to one of the following normal forms:

$$\begin{aligned} A_m & x_1^{m+1} + x_2^2 + \cdots + x_n^2 & (m \geq 1), \\ D_m & x_1^2 x_2 + x_2^{m-1} + x_3^2 + \cdots + x_n^2 & (m \geq 4), \\ E_6 & x_1^3 + x_2^4 + x_3^2 + \cdots + x_n^2, \\ E_7 & x_1^3 + x_1 x_2^3 + x_3^2 + \cdots + x_n^2, \\ E_8 & x_1^3 + x_2^5 + x_3^2 + \cdots + x_n^2. \end{aligned}$$

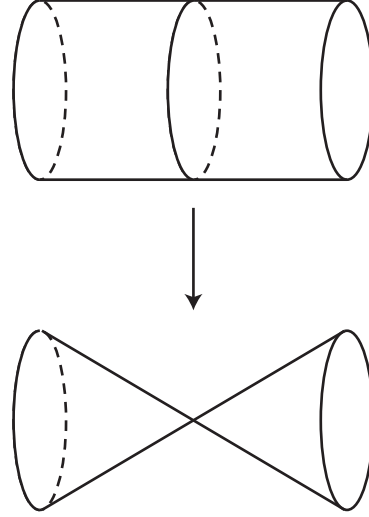


Figure 4 A resolution of the cone.

The names should bring to mind the CLASSIFICATION OF LIE GROUPS [III.50]. The connections are numerous but not easy to explain. When $n = 3$, these are also called *Du Val singularities* or *rational double points*.

Consider again the cone $z^2 = x^2 + y^2$. Earlier, we described a two-to-one parametrization of it. Here is another, and for many purposes better, parametrization over the real numbers.

In the (u, v, w) -space consider the smooth cylinder $u^2 + v^2 = 1$. The map $(u, v, w) \mapsto (uw, vw, w)$ maps the cylinder onto the cone (see figure 4). The map is one-to-one away from the vertex, the preimage of which is the circle $u^2 + v^2 = 1$ in the $(w = 0)$ -plane.

(Sharp-eyed readers will have noticed that this map is not so nice if we use complex numbers. In general, we want parametrizations that work both for real and complex numbers, but that would be quite a bit more complicated to describe.)

The advantage of the cylinder over the cone is that it does not have a singularity. Parametrizations of varieties in terms of smooth varieties are very useful, and there is a major result that tells us that they always exist, at least when the varieties are real or complex. (The corresponding result is still unknown for the finite geometries considered earlier.)

Resolution of singularities (Hironaka). For any variety X there is another smooth variety Y and a polynomially defined surjective map $\pi : Y \rightarrow X$ such that π is invertible at all smooth points of X .

(In the cone example above, one can take the whole cylinder, but the cylinder minus finitely many points in the collapsed circle would also work. In order to avoid such silly cases, we require π to be surjective in a very strong sense: if a sequence of smooth points $x_i \in X$ converges to a limit in X , then a subsequence of their preimages $\pi^{-1}(x_i)$ converges to a limit in Y .)

10 Classification of Curves

In order to get an idea of how the classification of algebraic varieties should proceed, let us look at hypersurfaces of degree d in n -space. These are given by a degree- d polynomial $f(x_1, \dots, x_n) = 0$. The set of all polynomials of degree at most d forms a vector space $V_{n,d}$. Thus hypersurfaces have two obvious discrete invariants, the dimension and the degree, and one can move between hypersurfaces of the same dimension and degree by varying the coefficients of f continuously. Moreover, the entire set $V_{n,d}$ is itself an algebraic variety. Our aim is to develop a similar understanding for all varieties, which can be done in two steps.

The first step is to define some integers, naturally attached to varieties, which stay the same if we change a variety continuously. Such integers are called *discrete invariants*. The simplest example is the dimension.

The second is to show that the set of all varieties with the same discrete invariant is parametrized by another algebraic variety, called the *MODULI SPACE* [IV.8]. Moreover, we would like the variety used for this parametrization to be chosen as economically as possible. We will look at this in more detail in the next section.

Let us see how it is accomplished for curves. Here there is only one more discrete invariant besides the dimension, known as the *genus* of the curve. This has many different definitions: one of the simplest is through topology. Let E be a smooth curve and let us look at its complex points. Locally, this set looks like \mathbb{C} , so it is a topological surface. After patching up some holes at infinity, we get a compact surface. Multiplication by $\sqrt{-1}$ gives an orientation, so basic topology tells us that we get a sphere with a certain number of handles attached (see DIFFERENTIAL TOPOLOGY [IV.7]). The genus of the curve is defined to be the number of these handles (that is, the genus of the corresponding surface). To see what this means in practice, let us look at some examples.

A line in 2-space is like the complex numbers, which can be viewed as a sphere minus a point. This sphere,

\mathbb{C} plus the point at infinity, is also called the *Riemann sphere*. So the genus is zero.

Next, we look at conics. Here it is better to use some projective geometry. Take any tangent of the conic and move this so that it becomes the line at infinity. Then we get a parabola, which, in suitable coordinates, is given by an equation $y = x^2$. The polynomial map $t \mapsto (t, t^2)$, with its inverse $(x, y) \mapsto x$, shows that this parabola is isomorphic to a line, so again has genus 0.

Cubics are quite a bit more complicated. A first warning is that $y = x^3$ is the wrong cubic to look at. It is smooth (and has genus 0) but it is singular at infinity. (The earlier expediency of keeping silent about projective geometry starts to bite us!) In any case, the correct thing to do is to choose the tangent line of the cubic at an inflection point and move that to infinity. After some computation we obtain a much-simplified equation $y^2 = f(x)$, where f has degree 3. What is the genus?

Consider the special case $y^2 = x(x-1)(x-2)$. We try to understand the two-to-one projection to the (complex) x -axis, but it is better to do this when the x -axis has already had the point at infinity added, so that it is the Riemann sphere. If we remove the interval $0 \leq x \leq 1$ and the half line $2 \leq x \leq +\infty$ from the Riemann sphere, then the function $y = \sqrt{x(x-1)(x-2)}$ has two branches. (This means that y takes two different values for each x , the positive and negative square roots of $x(x-1)(x-2)$, but if one moves x about, one can let y vary in a continuous way.) The sphere minus two slits is topologically like a cylinder, hence the complex cubic is glued together from two cylinders. So we get the torus and the genus is 1.

It turns out that a smooth plane curve of degree d has genus $\frac{1}{2}(d-1)(d-2)$, but I find this hard to see directly topologically.

It is a (probably hopeless) dream of algebraic geometers to give a similarly simple description of the discrete invariants for higher-dimensional varieties. Unfortunately, the topological invariants of the complex points are not good enough, and they probably mislead more than help.

As a further illustration of the approach to the classification of curves, here is a list of all curves of low genus.

Genus 0. There is only one curve of genus 0. As we saw, it can be realized as a line or as a conic in the plane.

Genus 1. Every curve of genus 1 is a plane cubic, and it can be given by an equation of the form $y^2 = f(x)$, where f has degree 3. Genus-1 curves are usually called **ELLIPTIC CURVES** [III.21], since they first appeared (in the guise of elliptic integrals) in connection with the arc length of ellipses. We look at these in more detail later.

Genus 2. Every curve of genus 2 can be given by an equation of the form $y^2 = f(x)$, where f has degree 5. (These curves are singular at infinity.) More generally, if f has degree $2g + 1$ or $2g + 2$, then the curve $y^2 = f(x)$ has genus g . For $g \geq 3$, such curves, called **hyperelliptic**, are rather special.

Genus 3. Every curve of genus 3 can be realized as a plane curve of degree 4 (or it is hyperelliptic).

Genus 4. Every curve of genus 4 can be presented as a space curve given by two equations of degrees 2 and 3 (or it is hyperelliptic).

It should be emphasized that hyperelliptic curves do not form a separate family. One can move continuously from any hyperelliptic curve to a general curve of the kind described above. This can be seen through more-complicated representations.

One can continue in this manner a bit longer, up to about genus 10, but no such explicit construction is possible when the genus is large.

11 Moduli Spaces

Let us go back to plane cubics, which we parametrized by the vector space $V_{2,3}$ of degree-3 polynomials in two variables. This is not very economical. For instance, $x^3 + 2y^3 + 1$ and $3x^3 + 6y^3 + 3$ are different polynomials, but define the same curve. Furthermore, there is not much reason to distinguish $x^3 + 2y^3 + 1$ from $2x^3 + y^3 + 1$, since they are obtained from each other by switching the two coordinate axes. More generally, as we have seen in the previous section, any cubic curve can be transformed into one given by an equation $y^2 = f(x)$, where $f = ax^3 + bx^2 + cx + d$.

This is better but not yet optimal, and there are two more steps to take. First, one can set the leading coefficient of f to be 1. Indeed, substitute $y = \sqrt{a}y_1$ and then divide the whole equation by a to get $y_1^2 = x^3 + \dots$. Second, we can make a substitution $x = ux_1 + v$ to get another elliptic curve with equation $y^2 = f(ux_1 + v) = f_1(x_1)$, where f_1 is easy to write down explicitly. One can see that these are the only coordinate changes that we can make without messing up the form $y^2 = (\text{cubic polynomial})$.

It is still not very clear what happens. To get a better answer, look at the three roots of f , so $f(x) = (x - r_1)(x - r_2)(x - r_3)$. (Again, complex numbers inevitably appear.) If we make the substitution $x \mapsto (r_2 - r_1)x + r_1$, we get a new polynomial $f_1(x)$, two of whose roots are 0 and 1. Thus our elliptic curve is transformed into $y^2 = x(x - 1)(x - \lambda)$. So instead of the four unknown coefficients of f , we are down to only one unknown, λ .

This form is still not completely unique. In our transformation we sent r_1, r_2 to 0, 1, but we could have used any two roots. For instance, we can substitute $x \mapsto 1 - x$, sending $\lambda \mapsto 1 - \lambda$, or $x \mapsto \lambda x$, sending $\lambda \mapsto \lambda^{-1}$. All together, the six values

$$\lambda, \frac{1}{\lambda}, 1 - \lambda, \frac{1}{1 - \lambda}, \frac{-\lambda}{1 - \lambda}, \frac{1 - \lambda}{-\lambda}$$

give “the same” elliptic curve. Most of the time these six values are different, but there may be coincidences. For instance, we get only three different values if $\lambda = -1$. This corresponds to the fact that the elliptic curve $y^2 = x(x - 1)(x + 1)$ has four symmetries: $(x, y) \mapsto (-x, \pm\sqrt{-1}y)$ and $(x, y) \mapsto (x, \pm y)$. (An unusual feature of elliptic curves is that they all have the second pair of symmetries. At $\lambda = 1$ we pick up 4/2 new symmetries, which corresponds to halving the number of different values above.)

The best way to think about it is to view this as an action of the symmetric group S_3 (the group of permutations of a three-element set) on the set $\mathbb{C} \setminus \{0, 1\}$.

It is not at all obvious that we have run out of tricks, but we have in fact reached the final result.

Moduli of elliptic curves. *The set of all elliptic curves is in a natural one-to-one correspondence with the points of the quotient orbifold $(\mathbb{C} \setminus \{0, 1\})/S_3$. The orbifold points correspond to the elliptic curves with extra automorphisms.*

This is the simplest illustration of a general phenomenon.

Moduli principle. *In most cases of interest, the set of all algebraic varieties with fixed discrete invariants is in a natural one-to-one correspondence with the points of an orbifold. The orbifold points correspond to the varieties with extra automorphisms.*

The moduli orbifold (also called the moduli space) of smooth curves of genus g is denoted by \mathcal{M}_g . These are among the most intensely studied orbifolds in algebraic geometry, especially since the recent discovery of their fundamental position in STRING THEORY [IV.17 §2] and MIRROR SYMMETRY [IV.16].

12 Effective Nullstellensatz

In order to show that there are still interesting elementary questions in algebraic geometry, let us try to decide when m given polynomials f_1, \dots, f_m have no common complex zero. The classical answer is given by the following result, which tells us that an obviously necessary condition is in fact sufficient.

Weak Nullstellensatz. *The polynomials f_1, \dots, f_m have no common complex zero if and only if there are polynomials g_1, \dots, g_m such that*

$$g_1 f_1 + \dots + g_m f_m = 1.$$

Let us now make a guess that we can find g_j with degree at most 100. We can then write

$$g_j = \sum_{i_1 + \dots + i_n \leq 100} a_{j,i_1,\dots,i_n} x_1^{i_1} \dots x_n^{i_n},$$

where the a_{j,i_1,\dots,i_n} are indeterminates. If we write $g_1 f_1 + \dots + g_m f_m$ as a polynomial in the variables x_1, \dots, x_n , then all the coefficients must vanish, save the constant term which must equal 1. Thus we get a system of *linear* equations in the indeterminates a_{j,i_1,\dots,i_n} . The solvability of systems of linear equations is well-known (with good computer implementations). Thus we can decide if there is a solution with $\deg g_j \leq 100$. Of course it is possible that 100 was too small a guess, and we may have to repeat the process with larger and larger degree bounds. Will this ever end? The answer is given by the following result, which was proved only recently.

Effective Nullstellensatz. *Let f_1, \dots, f_m be polynomials of degree less than or equal to d in n variables, where $d \geq 3$, $n \geq 2$. If they have no common zero, then $g_1 f_1 + \dots + g_m f_m = 1$ has a solution such that $\deg g_j \leq d^n - d$.*

For most systems, one can find solutions with $\deg g_j \leq (n-1)(d-1)$, but in general the upper bound $d^n - d$ cannot be improved.

As explained above, this provides a computational method for deciding whether or not a system of polynomial equations has a common solution. Unfortunately, this is rather useless in practice as we end up with exceedingly large linear systems. We still do not have a computationally effective and foolproof method.

13 So, What Is Algebraic Geometry?

To me algebraic geometry is a belief in the unity of geometry and algebra. The most exciting and profound

developments arise from the discovery of new connections. We have seen hints of some of these; many more were left unmentioned. Born with Cartesian coordinates, algebraic geometry is now intertwined with coding theory, number theory, computer-aided geometric design, and theoretical physics. Several of these connections have emerged in the last decade, and I hope to see many more in the future.

Further Reading

Most of the algebraic geometry literature is very technical. A notable exception is *Plane Algebraic Curves* (Birkhäuser, Boston, MA, 1986), by E. Brieskorn and H. Knörrer, which starts with a long overview of algebraic curves through arts and sciences since antiquity, with many nice pictures and reproductions. *A Scrapbook of Complex Curve Theory* (American Mathematical Society, Providence, RI, 2003), by C. H. Clemens, and *Complex Algebraic Curves* (Cambridge University Press, Cambridge, 1992), by F. Kirwan, also start at an easily accessible level, but then delve more quickly into advanced subjects.

The best introduction to the techniques of algebraic geometry is *Undergraduate Algebraic Geometry* (Cambridge University Press, Cambridge, 1988), by M. Reid. For those wishing for a general overview, *An Invitation to Algebraic Geometry* (Springer, New York, 2000), by K. E. Smith, L. Kahanpää, P. Kekäläinen, and W. Traves, is a good choice, while *Algebraic Geometry* (Springer, New York, 1995), by J. Harris, and *Basic Algebraic Geometry*, volumes I and II (Springer, New York, 1994), by I. R. Shafarevich, are suitable for more systematic readings.

IV.5 Arithmetic Geometry

Jordan S. Ellenberg

1 Diophantine Problems, Alone and in Teams

Our goal is to sketch some of the essential ideas of arithmetic geometry; we begin with a problem which, on the face of it, involves no geometry and only a bit of arithmetic.

Problem. Show that the equation

$$x^2 + y^2 = 7z^2 \tag{1}$$

has no solution in nonzero rational numbers x, y, z .

(Note that it is only in the coefficient 7 that (1) differs from the Pythagorean equation $x^2 + y^2 = z^2$, which we know has *infinitely* many solutions. It is a feature of

arithmetic geometry that modest changes of this kind can have drastic effects!)

Solution. Suppose x, y, z are rational numbers satisfying (1); we will derive from this a contradiction.

If n is the least common denominator of x, y, z , we can write

$$x = a/n, \quad y = b/n, \quad z = c/n$$

such that a, b, c , and n are integers. Our original equation (1) now becomes

$$\left(\frac{a}{n}\right)^2 + \left(\frac{b}{n}\right)^2 = 7\left(\frac{c}{n}\right)^2,$$

and multiplying through by n^2 one has

$$a^2 + b^2 = 7c^2. \quad (2)$$

If a, b , and c have a common factor m , then we can replace them by $a/m, b/m$, and c/m , and (2) still holds for these new numbers. We may therefore suppose that a, b , and c are integers with no common factor.

We now reduce the above equation modulo 7 (see MODULAR ARITHMETIC [III.60]). Denote by \bar{a} and \bar{b} the reductions of a and b modulo 7. The right-hand side of (2) is a multiple of 7, so it reduces to 0. We are left with

$$\bar{a}^2 + \bar{b}^2 = 0. \quad (3)$$

Now there are only seven possibilities for \bar{a} , and seven possibilities for \bar{b} . So the analysis of the solutions of (3) amounts to checking the forty-nine choices of \bar{a}, \bar{b} and seeing which ones satisfy the equation. A few minutes of calculation are enough to convince us that (3) is satisfied only if $\bar{a} = \bar{b} = 0$.

But saying that $\bar{a} = \bar{b} = 0$ is the same as saying that a and b are both multiples of 7. This being the case, a^2 and b^2 are both multiples of 49. It follows that their sum, $7c^2$, is a multiple of 49 as well. Therefore, c^2 is a multiple of 7, and this implies that c itself is a multiple of 7. In particular, a, b , and c share a common factor of 7. We have now arrived at the desired contradiction, since we chose a, b , and c to have no common factor. Thus, the hypothesized solution leads us to a contradiction, so we are forced to conclude that there is not, in fact, any solution to (1) consisting of nonzero rational numbers.¹

In general, the determination of rational solutions to a polynomial equation like (2) is called a *Diophantine problem*. We were able to dispose of (2) in a paragraph,

but that turns out to be the exception: in general, Diophantine problems can be extraordinarily difficult. For instance, we might modify the exponents in (2) and consider the equation

$$x^5 + y^5 = 7z^5. \quad (4)$$

I do not know whether (4) has any solutions in nonzero rational numbers or not; one can be sure, though, that determining the answer would be a substantial piece of work, and it is quite possible that the most powerful techniques available to us are insufficient to answer this simple question.

More generally, one can take an arbitrary commutative RING [III.83] R , and ask whether a certain polynomial equation has solutions in R . For instance, does (2) have a solution with x, y, z in the polynomial ring $\mathbb{C}[t]$? (The answer is yes. We leave it as an exercise to find some solutions.) We call the problem of solving a polynomial equation over R a *Diophantine problem over R* . The subject of arithmetic geometry has no precise boundary, but to a first approximation one may say that it concerns the solution of Diophantine problems over subrings of NUMBER FIELDS [III.65]. (To be honest, a problem is usually called Diophantine *only* when R is a subring of a number field. However, the more general definition suits our current purposes.)

With any particular equation like (2), one can associate *infinitely many* Diophantine problems, one for each commutative ring R . A central insight—in some sense the basic insight—of modern algebraic geometry is that this whole gigantic ensemble of problems can be treated as a single entity. This widening of scope reveals structure that is invisible if we consider each problem on its own. The aggregate we make of all these Diophantine problems is called a *scheme*. We will return to schemes later, and will try, without giving precise definitions, to convey some sense of what is meant by this not very suggestive term.

A word of apology: I will give only the barest sketch of the immense progress that has taken place in arithmetic geometry in recent decades—there is simply too much to cover in an article of the present scope. I have chosen instead to discuss at some length the idea of a scheme, assuming, I hope, minimal technical knowledge on the part of the reader. In the final section, I shall discuss some outstanding problems in arithmetic geometry with the help of the ideas developed in the body of the article. It must be conceded that the theory of schemes, developed by Grothendieck and his

1. Exercise: why does our argument not obtain a contradiction from the solution $x = y = z = 0$?

collaborators in the 1960s, belongs to algebraic geometry as a whole, and not to arithmetic geometry alone. I think, though, that in the arithmetic setting, the use of schemes, and the concomitant extension of geometric ideas to contexts that seem “nongeometric” at first glance, is particularly central.

2 Geometry without Geometry

Before we dive into the abstract theory of schemes, let us splash around a little longer among the polynomial equations of degree 2. Though it is not obvious from our discussion so far, the solution of Diophantine problems is properly classified as part of geometry. Our goal here will be to explain why this is so.

Suppose we consider the equation

$$x^2 + y^2 = 1. \quad (5)$$

One can ask: which values of $x, y \in \mathbb{Q}$ satisfy (5)? This problem has a flavor very different from that of the previous section. There we looked at an equation with *no* rational solutions. We shall see in a moment that (5), by contrast, has *infinitely* many rational solutions. The solutions $x = 0, y = 1$ and $x = \frac{3}{5}, y = -\frac{4}{5}$ are representative examples. (The four solutions $(\pm 1, 0)$ and $(0, \pm 1)$ are the ones that would be said, in the usual mathematical parlance, to be “staring you in the face.”)

Equation (5) is, of course, immediately recognizable as “the equation of a circle.” What, precisely, do we mean by that assertion? We mean that the set of pairs of real numbers (x, y) satisfying (5) forms a circle when plotted in the Cartesian plane.

So geometry, as usually construed, makes its entrance in the figure of the circle. Now suppose that we want to find more solutions to (5). One way to proceed is as follows. Let P be the point $(1, 0)$, and let L be a line through P of slope m . Then we have the following geometric fact.

- (G) The intersection of a line with a circle consists of either zero, one, or two points; the case of a single point occurs only when the line is tangent to the circle.

From (G) we conclude that, unless L is the tangent line to the circle at P , there is exactly one point other than P where the line intersects the circle. In order to find solutions (x, y) to (5), we must determine coordinates for this point. So suppose L is the line through $(1, 0)$ with slope m , which is to say it is the line L_m whose equation is $y = m(x - 1)$. Then in order to find the

x -coordinates of the points of intersection between L_m and the circle, we need to solve the simultaneous equations $y = m(x - 1)$ and $x^2 + y^2 = 1$; that is, we need to solve $x^2 + m^2(x - 1)^2 = 1$ or, equivalently,

$$(1 + m^2)x^2 - 2m^2x + (m^2 - 1) = 0. \quad (6)$$

Of course, (6) has the solution $x = 1$. How many other solutions are there? The geometric argument above leads us to believe that there is at most one solution to (6). Alternatively, we can use the following algebraic fact, which is analogous² to the geometric fact (G).

- (A) The equation $(1 + m^2)x^2 - 2m^2x + (m^2 - 1) = 0$ has either zero, one, or two solutions in x .

Of course, the conclusion of statement (A) holds for *any* nontrivial quadratic equation in x , not just (6); it is a consequence of the factor theorem.

In this case, it is not really necessary to appeal to any theorem; one can find by direct computation that the solutions of (6) are $x = 1$ and $x = (m^2 - 1)/(m^2 + 1)$. We conclude that the intersection between the unit circle and L_m consists of $(0, 1)$ and the point P_m with coordinates

$$\left(\frac{m^2 - 1}{m^2 + 1}, \frac{-2m}{m^2 + 1} \right). \quad (7)$$

Equation (7) establishes a correspondence $m \mapsto P_m$, which associates with each slope m a solution P_m to (5). What is more, since every point on the circle, other than $(1, 0)$ itself, is joined to $(1, 0)$ by a unique line, we find that we have established a one-to-one correspondence between slopes m and solutions, other than $(1, 0)$, to equation (5).

A very nice feature of this construction is that it allows us to construct solutions to (5) not only over \mathbb{R} but over smaller fields, like \mathbb{Q} : it is evident that, when m is rational, so are the coordinates of the solution yielded by (7). For example, taking $m = 2$ yields the solution $(\frac{3}{5}, -\frac{4}{5})$. In fact, not only does (7) show us that (5) admits infinitely many solutions over \mathbb{Q} , it also gives us an explicit way to *parametrize* the solutions in terms of a variable m . We leave it as an exercise to prove that the solutions of (5) over \mathbb{Q} , apart from $(1, 0)$, are in one-to-one correspondence with rational values of m . Alas, rare is the Diophantine problem whose solutions can be parametrized in this way! Still, polynomial equations like (5) with solutions that can be parametrized by one

2. Note that (A), unlike (G), contains no mention of tangency; that is because the notion of tangency is more subtle in the algebraic setting, as we will see in section 4 below.

or more variables play a special role in arithmetic geometry; they are called *rational varieties* and constitute by any measure the best-understood class of examples in the subject.

I want to draw your attention to one essential feature of this discussion. We relied on geometric intuition (e.g., our knowledge of facts like (G)) to give us ideas about how to construct solutions to (5). On the other hand, now that we have erected an algebraic justification for our construction, we can kick away our geometric intuition as needless scaffolding. It was a geometric fact about lines and circles that *suggested* to us that (6) should have only one solution other than $x = 1$. However, once one has had that thought, one can *prove* that there is at most one such solution by means of the purely algebraic statement (A), which involves no geometry whatsoever.

The fact that our argument can stand without any reference to geometry means that it can be applied in situations that might not, at first glance, seem geometric. For instance, suppose we wished to study solutions to (5) over the finite field \mathbb{F}_7 . Now this solution set would not seem rightfully to be called “a circle” at all—it is just a finite set of points! Nonetheless, our geometrically inspired argument still works perfectly. The possible values of m in \mathbb{F}_7 are 0, 1, 2, 3, 4, 5, 6, and the corresponding solutions P_m are $(-1, 0)$, $(0, -1)$, $(2, 2)$, $(5, 5)$, $(5, 2)$, $(2, 5)$, $(0, 1)$. These seven points, together with $(1, 0)$, form the whole solution set of (5) over \mathbb{F}_7 .

We have now started to reap the benefits of considering a whole bundle of Diophantine problems at once; in order to find the solutions to (5) over \mathbb{F}_7 , we used a method that was inspired by the problem of finding solutions to (5) over \mathbb{R} . Similarly, in general, methods suggested by geometry can help us solve Diophantine problems. And these methods, once translated into purely algebraic form, still apply in situations that do not appear to be geometric.

We must now open our minds to the possibility that the purely algebraic appearance of certain equations is deceptive. Perhaps there could be a sense of “geometry” that was general enough to include entities like the solution set of (5) over \mathbb{F}_7 , and in which this particular example had every right to be called a “circle.” And why not? It has properties a circle has: most importantly for us, it has either zero, one, or two intersection points with any line. Of course, there are features of “circle-ness” which this set of points lacks: infinitude, continuity, roundness, etc. But these latter qualities turn out to be inessential when we are doing arithmetic geometry.

From our viewpoint the set of solutions of (5) over \mathbb{F}_7 has every right to be called the unit circle.

To sum up, you might think of the modern point of view as an upending of the traditional story of Cartesian space. There, we have geometric objects (curves, lines, points, surfaces) and we ask questions such as, “What is the equation of this curve?” or “What are the coordinates of that point?” The underlying object is the geometric one, and the algebra is there to tell us about its properties. For us, the situation is exactly reversed: the underlying object is the *equation*, and the various geometric properties of solution sets of the equation are merely tools that tell us about the equation’s algebraic properties. For an arithmetic geometer, “the unit circle” is the equation $x^2 + y^2 = 1$. And the round thing on the page? That is just a *picture* of the solutions to the equation over \mathbb{R} . It is a distinction that makes a remarkable difference.

3 From Varieties to Rings to Schemes

In this section, we will attempt to give a clearer answer to the question, “What is a scheme?” Instead of trying to lay out a precise definition—which requires more algebraic apparatus than would fit comfortably here—we will approach the question by means of an analogy.

3.1 Adjectives and Qualities

So let us think about adjectives. Any adjective, such as “yellow” for instance, picks out a set of nouns to which the adjective applies. For each adjective A , we might call this set of nouns $\Gamma(A)$. For instance, $\Gamma(\text{“yellow”})$ is an infinite set that might look like $\{\text{lemon, school bus, banana, sun, } \dots\}$.³ And anyone would agree that $\Gamma(A)$ is an important thing to know about A .

Now suppose that, moved by a desire for lexical parsimony, a theoretician among us suggested that adjectives could in fact be dispensed with entirely. If, instead of A , we spoke only of $\Gamma(A)$, we could get by with a grammatical theory involving only nouns.

Is this a good idea? Well, there are certainly some obvious ways that things could go wrong. For instance, what if lots of different adjectives were sent to the same set of nouns? Then our new viewpoint would be less precise than the old one. But it certainly seems that if two adjectives apply to *exactly* the same set of nouns,

3. Of course, in real life, there are nouns whose relationship with “yellow” is not so clear-cut, but since our goal is to make this look like mathematics, let us pretend that every object in the world is either definitively yellow or definitively not yellow.

then it is fair to say that the adjectives are the same, or at least synonymous.

What about relationships between adjectives? For instance, we can ask of two adjectives whether one is *stronger* than another, in the way that “gigantic” is stronger than “large.” Is this relationship between adjectives still visible on the level of sets of nouns? The answer is yes: it seems fair to say that A is “stronger than” B precisely when $\Gamma(A)$ is a subset of $\Gamma(B)$. In other words, what it means to say that “gigantic” is stronger than “large” is that all gigantic things are large, though some large things may not be gigantic.

So far, so good. We have paid a price in technical difficulty: it is much more cumbersome to speak of infinite sets of nouns than it was to use simple, familiar adjectives. But we have gained something, too: the opportunity for generalization. Our theoretician—whom we may now call a “set-theoretic grammarian”—observes that there is, perhaps, nothing special about the sets of nouns that happen to be of the form $\Gamma(A)$ for some already known adjective A . Why not take a conceptual leap and *redefine* the word “adjective” to mean “a set of nouns”? To avoid confusion with the usual meaning of “adjective,” the theoretician might even use a new term, like “quality,” to refer to his new objects of study.

Now we have a whole new world of qualities to play with. For example, there is a quality {“school bus”, “sun”} which is stronger than “yellow,” and a quality {“sun”} (not the same thing as the *noun* “sun”!) which is stronger than the qualities “yellow,” “gigantic,” “large,” and {“school bus”, “sun”}.

I may not have convinced you that, on balance, this reconception of the notion of “adjective” is a good idea. In fact, it probably is not, which is why set-theoretic grammar is not a going concern. The corresponding story in algebraic geometry, however, is quite a different matter.

3.2 Coordinate Rings

A warning: the next couple of sections will be difficult going for those not familiar with rings and ideals—such readers can either skip to section 4, or try to follow the discussion after reading RINGS, IDEALS, AND MODULES [III.83] (see also ALGEBRAIC NUMBERS [IV.1]).

Let us recall that a *complex affine variety* (from now on, just “variety”) is the set of solutions over \mathbb{C} to some finite set of polynomial equations. For instance, one variety V we could define is the set of points (x, y)

in \mathbb{C}^2 satisfying our favorite equation

$$x^2 + y^2 = 1. \quad (8)$$

Then V is what we called in the previous section “the unit circle,” though in fact the shape of the set of complex solutions of (8) is a sphere with two points removed. (This is not supposed to be obvious.) It is a question of general interest, given some variety X , to understand the ring of polynomial functions that take points on X to complex numbers. This ring is called the *coordinate ring* of X , and is denoted $\Gamma(X)$.

Certainly, given any polynomial in x and y , we can regard it as a function defined on our particular variety V . So is the coordinate ring of V just the polynomial ring $\mathbb{C}[x, y]$? Not quite. Consider, for instance, the function $f = 2x^2 + 2y^2 + 5$. If we evaluate this function at various points on V ,

$$f(0, 1) = 7, \quad f(1, 0) = 7,$$

$$f(1/\sqrt{2}, 1/\sqrt{2}) = 7, \quad f(i, \sqrt{2}) = 7, \quad \dots,$$

we notice that f keeps taking the same value; indeed, since $x^2 + y^2 = 1$ for all $(x, y) \in V$, we see that $f = 2(x^2 + y^2) + 5$ takes the value 7 at *every* point on V . So $2x^2 + 2y^2 + 5$ and 7 are just different names for the same function on V .

So $\Gamma(V)$ is smaller than $\mathbb{C}[x, y]$; it is the ring obtained from $\mathbb{C}[x, y]$ by declaring two polynomials f and g to be the same function whenever they take the same value at every point of V . (More formally, we are defining an EQUIVALENCE RELATION [I.2 §2.3] on the set of complex polynomials in two variables.) It turns out that f and g have this property precisely when their difference is a multiple of $x^2 + y^2 - 1$. Thus, the ring of polynomial functions on V is the quotient of $\mathbb{C}[x, y]$ by the ideal generated by $x^2 + y^2 - 1$. This ring is denoted by $\mathbb{C}[x, y]/(x^2 + y^2 - 1)$.

We have shown how to attach a ring of functions to any variety. It is not hard to show that, if X and Y are two varieties, and if their coordinate rings $\Gamma(X)$ and $\Gamma(Y)$ are ISOMORPHIC [I.3 §4.1], then X and Y are in a sense the “same” variety. It is a short step from this observation to the idea of abandoning the study of varieties entirely in favor of the study of rings. Of course, we are here in the position of the set-theoretic grammarian in the parable above, with “variety” playing the part of “adjective” and “coordinate ring” the part of “set of nouns.”

Happily, we can recover the geometric properties of a variety from the algebraic properties of its coordinate ring; if this were not the case, the coordinate ring would

PUP: Tim prefers the current wording to that suggested by the proofreader. OK?

PUP: Tim thinks the two sentences in a row that start with ‘So’ are OK as it fits with the author’s conversational style. OK?

not be such a useful object! The relationship between geometry and algebra is a long story—and much of it belongs to algebraic geometry in general, not arithmetic geometry in particular—but to give the flavor, let us discuss some examples.

A straightforward geometric property of a variety is *irreducibility*. We say a variety X is *reducible* if X can be expressed as the union of two varieties X_1 and X_2 , neither of which is the whole of X . For example, the variety

$$x^2 = y^2 \quad (9)$$

in \mathbb{C}^2 is the union of the lines $x = y$ and $x = -y$. A variety is called *irreducible* if it is not reducible. All varieties are thus built up from irreducible varieties: the relationship between irreducible varieties and general varieties is rather like the relationship between prime numbers and general positive integers.

Moving from geometry to algebra, we recall that a ring R is called an *integral domain* if, whenever f, g are nonzero elements of R , their product fg is also nonzero; the ring $\mathbb{C}[x, y]$ is a good example.

Fact. A variety X is irreducible if and only if $\Gamma(X)$ is an integral domain.

Experts will note that we are glossing over issues of “reducedness” here.

We will not prove this fact, but the following example is illustrative: consider the two functions $f = x - y$ and $g = x + y$ on the variety X defined by (9). Neither of these functions is the zero function; note, for instance, that $f(1, -1)$ is nonzero, as is $g(1, 1)$. Their product, however, is $x^2 - y^2$, which is equal to zero on X ; so $\Gamma(X)$ is not an integral domain. Notice that the functions f and g that we chose are closely related to the decomposition of X as the union of two smaller varieties.

Another crucial geometric notion is that of functions from one variety to another. (It is common practice to call such functions “maps” or “morphisms”; we will use the three words interchangeably.) For instance, suppose that W is the variety in \mathbb{C}^3 determined by the equation $xyz = 1$. Then the map $F : \mathbb{C}^3 \rightarrow \mathbb{C}^2$ defined by

$$F(x, y, z) = \left(\frac{1}{2}(x + yz), \frac{1}{2i}(x - yz) \right)$$

maps points of W to points of V .

It turns out that knowing the coordinate rings of varieties makes it very easy to see the maps between the varieties. We merely observe that if $G : V_1 \rightarrow V_2$ is a map between varieties V_1 and V_2 , and if f is a

polynomial function on V_2 , then we have a polynomial function on V_1 that sends every point v to $f(G(v))$. This function on V_1 is denoted by $G^*(f)$. For example, if f is the function $x + y$ on V , and F is the map above, $F^*(f) = \frac{1}{2}(x + yz) + (1/2i)(x - yz)$. It is easy to check that G^* is a \mathbb{C} -algebra homomorphism (that is, a homomorphism of rings that sends each element of \mathbb{C} to itself) from $\Gamma(V_2)$ to $\Gamma(V_1)$. What is more, one has the following theorem.

Fact. For any pair of varieties V, W , the correspondence sending G to G^* is a bijection between the polynomial functions sending W to V and the \mathbb{C} -algebra homomorphisms from $\Gamma(V)$ to $\Gamma(W)$.

You would not be far off in thinking of the statement “there is an injective map from V to W ” as analogous to “quality A is stronger than quality B .”

The move to transform geometry into algebra is not something one undertakes out of sheer love of abstraction, or hatred of geometry. Instead, it is part of the universal mathematical instinct to unify seemingly disparate theories. I cannot put it any better than Dieudonné (1985) does in his *History of Algebraic Geometry*:

... from [the 1882 memoirs of] Kronecker and Dedekind-Weber dates the awareness of the profound analogies between algebraic geometry and the theory of algebraic numbers, which originated at the same time. Moreover, this conception of algebraic geometry is the most simple and most clear for us, trained as we are in the wielding of “abstract” algebraic notions: rings, ideals, modules, etc. But it is precisely this “abstract” character that repulsed most contemporaries, disconcerted as they were by not being able to recover the corresponding geometric notions easily. Thus the influence of the algebraic school remained very weak up until 1920. ... It certainly seems that Kronecker was the first to dream of one vast algebraic-geometric construction comprising these two theories at once; this dream has begun to be realized only recently, in our era, with the theory of schemes.

Let us therefore move on to schemes.

3.3 Schemes

We have seen that each variety X gives rise to a ring $\Gamma(X)$, and furthermore that the algebraic study of these rings can stand in for the geometric study of varieties. But just as not every set of nouns corresponds to an adjective, not every ring arises as the coordinate ring of a variety. For example, the ring \mathbb{Z} of integers is not

the coordinate ring of a variety, as we can see by the following argument: for every complex number a and every variety V , the constant function a is a function on V , and therefore $\mathbb{C} \subset \Gamma(V)$ for every variety V . Since \mathbb{Z} does not contain \mathbb{C} as a subring, it is not the coordinate ring of any variety.

Now we are ready to imitate the set-theoretic grammarian's coup de grâce. We know that some, but not all, rings arise from geometric objects (varieties); and we know that the geometry of these varieties is described by algebraic properties of these special rings. Why not, then, just consider *every* ring R to be a “geometric object” whose geometry is determined by algebraic properties of R ? The grammarian needed to invent a new word, “quality,” to describe his generalized adjectives; we are in the same position with our rings-that-are-not-coordinate-rings; we will call them *schemes*.

So, after all this work, the definition of scheme is rather prosaic—schemes are rings! (In fact, we are hiding some technicalities; it is correct to say that *affine schemes* are rings. Restricting our attention to affine schemes will not interfere with the phenomena that we are aiming to explain.) More interesting is to ask how we can carry out the task whose difficulty “disconcerted” the early algebraic geometers—how can we identify “geometric” features of arbitrary rings?

For instance, if R is supposed to be an arbitrary geometric object, it ought to have “points.” But what are the “points” of a ring? Clearly we cannot mean by this the *elements* of the ring; for in the case $R = \Gamma(X)$, the elements of R are *functions* on X , not points on X . What we need, given a point p on X , is some entity attached to the ring R that corresponds to p .

The key observation is that we can think of p as a map from $\Gamma(X)$ to \mathbb{C} : given a function f from $\Gamma(X)$ we map it to the complex number $f(p)$. This map is a homomorphism, called the *evaluation homomorphism at p* . Since points on X give us homomorphisms on $\Gamma(X)$, a natural way to define the word “point” for the ring $R = \Gamma(X)$, without using geometry, is to say that a “point” is a homomorphism from R to \mathbb{C} . It turns out that the kernel of such a homomorphism is a prime ideal. Moreover, with the exception of the zero ideal, every prime ideal of R arises from a point p of X . So a very concise way to describe the points of X might be to say that they are the nonzero prime ideals of R .

The definition we have arrived at makes sense for *all* rings R , and not just those of the form $R = \Gamma(X)$. So we might define the “points” of a ring R to be its prime ideals. (Considering all prime ideals, rather than

only the nonzero ones, turns out to be a wiser technical choice.) The set of prime ideals of R is given the name $\text{Spec } R$, and it is $\text{Spec } R$ that we call the *scheme associated with R* . (More precisely, $\text{Spec } R$ is defined to be a “locally ringed topological space” whose points are the prime ideals of R , but we will not need the full power of this definition for our discussion here.)

We are now in a position to elucidate our claim, made in the first section, that a scheme incorporates into one package Diophantine problems over many different rings. Suppose, for instance, that R is the ring $\mathbb{Z}[x, y]/(x^2 + y^2 - 1)$. We are going to catalog the homomorphisms $f : R \rightarrow \mathbb{Z}$. To specify f , I merely have to tell you the values of $f(x)$ and $f(y)$ in \mathbb{Z} . But I cannot choose these values arbitrarily: since $x^2 + y^2 - 1 = 0$ in R , it must be the case that

$$f(x)^2 + f(y)^2 - 1 = 0$$

in \mathbb{Z} . In other words, the pair $(f(x), f(y))$ constitutes a solution over \mathbb{Z} to the Diophantine equation $x^2 + y^2 = 1$. What is more, the same argument shows that, for *any* ring S , a homomorphism $f : R \rightarrow S$ yields a solution over S to $x^2 + y^2 = 1$, and vice versa. In summary,

for each S , there is a one-to-one correspondence between the set of ring homomorphisms from R to S , and solutions over S to $x^2 + y^2 = 1$.

This behavior is what we have in mind when we say that the ring R “packages” information about Diophantine equations over different rings.

It turns out, just as one might hope, that every interesting geometric property of varieties can be computed by means of the coordinate ring, which means it can be defined, not only for varieties, but for general schemes. We have already seen, for instance, that a variety X is irreducible if and only if $\Gamma(X)$ is an integral domain. Thus, we say in general that a scheme $\text{Spec } R$ is irreducible if and only if R is an integral domain (or, more precisely, if the quotient of R by its nilradical is an integral domain). One can speak of the connectedness of a scheme, its dimension, whether it is smooth, and so forth. All these geometric properties turn out, like irreducibility, to have purely algebraic descriptions. In fact, to the arithmetic geometer's way of thinking, all these *are*, at bottom, algebraic properties.

3.4 Example: $\text{Spec } \mathbb{Z}$, the Number Line

The first ring we encounter in our mathematical education—and the ring that is the ultimate subject of number theory—is \mathbb{Z} , the ring of integers. How does it

fit into our picture? The scheme $\text{Spec } \mathbb{Z}$ has as its points the set of prime ideals of \mathbb{Z} , which come in two flavors: there are the principal ideals (p) , with p a prime number; and there is the zero ideal. (The fact that these are the only prime ideals of \mathbb{Z} is not a triviality; it can be derived from the EUCLIDEAN ALGORITHM [III.22].)

We are supposed to think of \mathbb{Z} as the ring of “functions” on $\text{Spec } \mathbb{Z}$. How can an integer be a function? Well, I merely need to tell you how to evaluate an integer n at a point of $\text{Spec } \mathbb{Z}$. If the point is a nonzero prime ideal (p) , then the evaluation homomorphism at (p) is precisely the homomorphism whose kernel is (p) ; so the value of n at (p) is just the reduction of n modulo p . At the point (0) , the evaluation homomorphism is the identity map $\mathbb{Z} \rightarrow \mathbb{Z}$; so the value of n at (0) is just n .

4 How Many Points Does a Circle Have?

We now return to the method of section 2, paying particular attention to the case where the equation $x^2 + y^2 = 1$ is considered over a finite field \mathbb{F}_p .

Let us write V for the scheme of solutions of $x^2 + y^2 = 1$. For any ring R , we will denote by $V(R)$ the set of solutions of $x^2 + y^2 = 1$.

If R is a finite field \mathbb{F}_p , the set $V(\mathbb{F}_p)$ is a subset of \mathbb{F}_p^2 . In particular, it is a *finite* set. So it is natural to wonder how large this set is: in other words, how many points does a circle have?

In section 2, guided by our geometric intuition, we observed that, for every $m \in \mathbb{Q}$, the point

$$P_m = \left(\frac{m^2 - 1}{m^2 + 1}, \frac{-2m}{m^2 + 1} \right)$$

lies on V .

The algebraic computation showing that P_m satisfies the equation $x^2 + y^2 = 1$ is no different over a finite field. So we might be inclined to think that $V(\mathbb{F}_p)$ consists of $p + 1$ points: namely, the points P_m for each $m \in \mathbb{F}_p$, together with $(1, 0)$.

But this is not right: for instance, when $p = 5$ it is easy to check that the four points $(0, 1)$, $(0, -1)$, $(1, 0)$, $(-1, 0)$ make up all of $V(\mathbb{F}_5)$. Computing P_m for various m , we quickly discover the problem; when m is 2 or 3, the formula for P_m does not make sense, because the denominator $m^2 + 1$ is zero! This is a wrinkle we did not see over \mathbb{Q} , where $m^2 + 1$ was always positive.

What is the geometric story here? Consider the intersection of the line L_2 , that is, the line $y = 2(x - 1)$, with V . If (x, y) belongs to this intersection, then we

have

$$\begin{aligned} x^2 + (2(x - 1))^2 &= 1, \\ 5x^2 - 8x + 3 &= 0. \end{aligned}$$

Since $5 = 0$ and $8 = 3$ in \mathbb{F}_5 , the above equation can be written as $3 - 3x = 0$; in other words, $x = 1$, which in turn implies that $y = 0$. In other words, the line L_2 intersects the circle V at only one point!

We are left with two possibilities, both disturbing to our geometric intuition. We might declare that L_2 is tangent to V ; but this means that V would have multiple tangents at $(1, 0)$, since the vertical line $x = 1$ should surely still be considered a tangent. The alternative is to declare that L_2 is *not* tangent to V ; but then we are in the equally unsavory situation of having a line which, while not tangent to the circle V , intersects it at only one point. You are now beginning to see why I did not include an algebraic definition of “tangent” in statement (A) above!

This quandary illustrates the nature of arithmetic geometry nicely. When we move into novel contexts, like geometry over \mathbb{F}_p , some features stay fixed (such as “a line intersects a circle in at most two points”), while others have to be discarded (such as “there exists exactly one line, which we may call the tangent line to the circle at $(1, 0)$, that intersects the circle at $(1, 0)$ and no other point”⁴).

Notwithstanding these subtleties, we are now ready to compute the number of points in $V(\mathbb{F}_p)$. First of all, when $p = 2$ one can check directly that $(0, 1)$ and $(1, 0)$ are the only two points in $V(\mathbb{F}_2)$. (Another common refrain in arithmetic geometry is that fields of characteristic 2 often impose technical annoyances, and are best dealt with separately.) Having treated this case, we assume for the rest of this section that p is odd. It follows from basic number theory that the equation $m^2 + 1 = 0$ has a solution in \mathbb{F}_p if and only if $p \equiv 1 \pmod{4}$, in which case there are exactly two such m . So, if $p \equiv 3 \pmod{4}$, then every line L_m intersects the circle at a point other than $(1, 0)$, and we have $p + 1$ points in all. If $p \equiv 1 \pmod{4}$, there are two choices of m for which L_m intersects V only at $(1, 0)$; eliminating these two choices of m yields a total of $p - 1$ points in $V(\mathbb{F}_p)$.

We conclude that $|V(\mathbb{F}_p)|$ is equal to 2 when $p = 2$, to $p - 1$ when $p \equiv 1 \pmod{4}$, and to $p + 1$ when $p \equiv 3 \pmod{4}$. The interested reader will find the following

4. In this case, the right attitude to adopt is that L_2 is not tangent to V , but that there are certain nontangent lines that intersect the circle at a single point.

exercises useful: how many solutions are there to $x^2 + 3y^2 = 1$ over \mathbb{F}_p ? What about $x^2 + y^2 = 0$?

More generally, let X be the scheme of solutions of *any* system of equations

$$F_1(x_1, \dots, x_n) = 0, F_2(x_1, \dots, x_n) = 0, \dots, \quad (10)$$

where the F_i are polynomials with integral coefficients. Then one can associate with F a list of integers $N_2(X), N_3(X), N_5(X), \dots$, where $N_p(X)$ is the number of solutions to (10) with $x_1, \dots, x_n \in \mathbb{F}_p$. This list of integers turns out to contain a surprising amount of geometric information about the scheme X ; even for the simplest schemes, the analysis of these lists is a deep problem of intense current interest, as we will see in the next section.

5 Some Problems in Classical and Contemporary Arithmetic Geometry

In this section I will try to give an impression of a few of arithmetic geometry's great successes, and to gesture at some problems of current interest for researchers in the area.

A word of warning is in order. In what follows, I will be trying to give brief and nontechnical descriptions of some mathematics of extreme depth and complexity. Consequently, I will feel very free to oversimplify. I will try to avoid making assertions that are actually false, but I will often use definitions (like that of the L -function attached to an elliptic curve) that do not exactly agree with those in the literature.

5.1 From Fermat to Birch-Swinnerton-Dyer

The world is not lacking in expositions of the proof of FERMAT'S LAST THEOREM [V.12] and I will not attempt to give another one here, although it is without question the most notable contemporary achievement in arithmetic geometry. (Here I am using the mathematician's sense of "contemporary," which, as the old joke goes, means "theorems proved since I entered graduate school." The shorthand for "theorems proved before I entered graduate school" is "classical.") I will content myself with making some comments about the structure of the proof, emphasizing connections with the parts of arithmetic geometry we have discussed above.

Fermat's last theorem (rightly called "Fermat's conjecture," since it is almost impossible to imagine that FERMAT [VI.12] proved it) asserts that the equation

$$A^\ell + B^\ell = C^\ell, \quad (11)$$

where ℓ is an odd prime, has no solutions in positive integers A, B, C .

The proof uses the crucial idea, introduced independently by Frey and Hellegouarch, of associating with any solution (A, B, C) of (11) a certain variety $X_{A,B}$, namely the curve described by the equation

$$y^2 = x(x - A^\ell)(x + B^\ell).$$

What can we say about $N_p(X_{A,B})$? We begin with a simple heuristic. There are p choices for x in \mathbb{F}_p . For each choice of x , there are either zero, one, or two choices for y , depending on whether $x(x - A^\ell)(x + B^\ell)$ is a quadratic nonresidue, zero, or a quadratic residue in \mathbb{F}_p . Since there are equally many quadratic residues and nonresidues in \mathbb{F}_p , we might guess that those two cases arise equally often. If so, there would on average be one choice of y for each of the p choices of x , which inclines us to make the estimate $N_p(X_{A,B}) \sim p$. Define a_p to be the error in this estimate: $a_p = p - N_p(X_{A,B})$. It is worth remembering that when X was the scheme attached to $x^2 + y^2 = 1$, the behavior of $p - N_p(X)$ was very regular; in particular, this quantity took the value 1 at primes congruent to 1 mod 4 and -1 at primes congruent to 3 mod 4. (We note, in particular, that the heuristic estimate $N_p(X) \sim p$ is quite good in this case.) Might one hope that a_p displays the same kind of regularity?

In fact, the behavior of the a_p is very *irregular*, as a famous theorem of Mazur shows; not only do the a_p fail to vary periodically, even their reductions modulo various primes are irregular!

Fact (Mazur). Suppose that ℓ is a prime greater than 3, and let b be a positive integer. It is not the case that a_p takes the same value (mod ℓ) for all primes p congruent to 1 (mod b).⁵

On the other hand—if I may compress a 200-page paper into a slogan—Wiles proved that, when A, B, C is a solution to (11), the reductions mod ℓ of the a_p *necessarily* behaved periodically, contradicting Mazur's theorem when $\ell > 3$. The case $\ell = 3$ is an old theorem of EULER [VI.19]. This completes the proof of Fermat's conjecture, and, I hope, bolsters our assertion that the careful study of the values $N_p(X)$ is an interesting way to study a variety X !

5. The theorem proved by Mazur is stated by him in a very different and much more general way: he proves that certain *modular curves* do not possess any rational points. This implies that a version of the fact above is true, not only for $X_{A,B}$, but for *any* equation of the form $y^2 = f(x)$, where f is a cubic polynomial without repeated roots. We will leave it to the other able treatments of Fermat to develop that point of view.

But the story does not end with Fermat. In general, if $f(x)$ is a cubic polynomial with coefficients in \mathbb{Z} and no repeated roots, the curve E defined by the equation

$$y^2 = f(x) \quad (12)$$

is called an **ELLIPTIC CURVE** [III.21] (note well that an elliptic curve is not an ellipse). The study of rational points on elliptic curves (that is, pairs of rational numbers satisfying (12)) has been occupying arithmetic geometers since before our subject existed as such; a decent treatment of the story would fill a book, as indeed it does fill the book of Silverman and Tate (1992). We can define $a_p(E)$ to be $p - N_p(E)$ as above. First of all, if our heuristic $N_p(E) \sim p$ is a good estimate, we might expect that $a_p(E)$ is small compared with p ; and, in fact, a theorem of Hasse from the 1930s shows that $a_p(E) \leq 2\sqrt{p}$ for all but finitely many p .

It turns out that some elliptic curves have infinitely many rational points, and some only finitely many. One might expect that an elliptic curve with many points over \mathbb{Q} would tend to have more points over finite fields as well, since the coordinates of a rational point can be reduced mod p to yield a point over the finite field \mathbb{F}_p . Conversely, one might imagine that, by knowing the list of numbers a_p , one could draw conclusions about the points of E over \mathbb{Q} .

In order to draw such conclusions, one needs a nice way to package the information of the infinite list of integers a_p . Such a package is given by the **L-FUNCTION** [III.49] of the elliptic curve, defined to be the following function of a variable s :

$$L(E, s) = \prod_p' (1 - a_p p^{-s} + p^{1-2s})^{-1}. \quad (13)$$

The notation \prod' means that this product is evaluated over all primes apart from a finite set, which is easy to determine from the polynomial f . (As is often the case, we are oversimplifying; what I have written here differs in some irrelevant-to-us respects from what is usually called $L(E, s)$ in the literature.) It is not hard to check that (13) is a convergent product when s is a real number greater than $\frac{3}{2}$. Not much deeper is the fact that the right-hand side of (13) is well-defined when s is a complex number whose real part exceeds $\frac{3}{2}$. What is much deeper—following from the theorem of Wiles, together with later theorems of Breuil, Conrad, Diamond, and Taylor—is that we can extend $L(E, s)$ to a **HOLOMORPHIC FUNCTION** [I.3 §5.6] defined for every complex number s .

A heuristic argument might suggest the following relationship between the values of $N_p(E)$ and the

value of $L(E, 1)$. If the a_p are typically negative (corresponding to the $N_p(E)$ typically being greater than p) the terms in the infinite product tend to be smaller than 1; when the a_p are positive, the terms in the product tend to be larger than 1. In particular, one might expect the value of $L(E, 1)$ to be closer to 0 when E has many rational points. Of course, this heuristic should be taken with a healthy pinch of salt, given that $L(E, 1)$ is not in fact defined by the infinite product on the right-hand side of (13)! Nonetheless, the **BIRCH-SWINNERTON-DYER CONJECTURE** [V.4], which makes precise the heuristic prediction above, is widely believed, and supported by many partial results and numerical experiments. We do not have the space here to state the conjecture in full generality. However, the following conjecture would follow from Birch-Swinnerton-Dyer.

Conjecture. The elliptic curve E has infinitely many points over \mathbb{Q} if and only if $L(E, 1) = 0$.

Kolyvagin proved one direction of this conjecture in 1988: that E has finitely many rational points if $L(E, 1) \neq 0$. (To be precise, he proved a theorem that yields the assertion here once combined with the later theorems of Wiles and others.) It follows from a theorem of Gross and Zagier that E has infinitely many rational points if $L(E, s)$ has a *simple* zero at $s = 1$. That more or less sums up our present knowledge about the relationship between L -functions and rational points on elliptic curves. This lack of knowledge has not, however, prevented us from constructing a complex of ever more rarefied conjectures in the same vein, of which the Birch-Swinnerton-Dyer conjecture is only a tiny and relatively down-to-earth sliver.

Before we leave the subject of counting points behind, we will pause and point out one more beautiful result: the theorem of **ANDRÉ WEIL** [VI.93] bounding the number of points on a curve over a finite field. (Because we have not introduced projective geometry, we will satisfy ourselves with a somewhat less beautiful formulation than the usual one.) Let $F(x, y)$ be an irreducible polynomial in two variables, and let X be the scheme of solutions of $F(x, y) = 0$. Then the complex points of X define a certain subset of \mathbb{C}^2 , which we call an *algebraic curve*. Since X is obtained by imposing one polynomial condition on the points of \mathbb{C}^2 , we expect that X has complex dimension 1, which is to say it has real dimension 2. Topologically speaking, $X(\mathbb{C})$ is, therefore, a surface. It turns out that, for almost all choices of F , the surface $X(\mathbb{C})$ will have the topology

T&T note: check style of CR later.

of a “ g -holed doughnut” with d points removed, for some nonnegative integers g and d . In this case we say that X is a *curve of genus g* .

In section 2 we saw that the behavior of schemes over finite fields seemed to “remember” facts arising from our geometric intuition over \mathbb{R} and \mathbb{C} : our example there was the fact that circles and lines intersect in at most two points.

The theorem of Weil reveals a similar, though much deeper, phenomenon.

Fact. Suppose the scheme X of solutions of $F(x, y)$ is a curve of genus g . Then, for all but finitely many primes p , the number of points of X over \mathbb{F}_p is at most $p + 1 + 2g\sqrt{p}$ and at least $p + 1 - 2g\sqrt{p} - d$.

Weil’s theorem illustrates the startlingly close bonds between geometry and arithmetic. The more complicated the topology of $X(\mathbb{C})$, the further the number of \mathbb{F}_p -points can vary from the “expected” answer of p . What is more, it turns out that knowing the size of the set $X(\mathbb{F}_q)$ for every finite field \mathbb{F}_q allows us to determine the genus of X . In other words, the *finite sets of points* $X(\mathbb{F}_q)$ somehow “remember” the topology of the space of complex points $X(\mathbb{C})$! In modern language, we say that there is a theory applying to general schemes, called *étale cohomology*, which mimics the theory of cohomology applying to the topology of varieties over \mathbb{C} .

Let us return for a moment to our favorite curve, by taking the polynomial $F(x, y) = x^2 + y^2 - 1$. In this case, it turns out that $X(\mathbb{C})$ has $g = 0$ and $d = 2$: our previous result that $X(\mathbb{F}_p)$ contains either $p + 1$ or $p - 1$ points therefore conforms exactly with the Weil bounds. We also remark that elliptic curves always have genus 1; so the theorem of Hasse alluded to above is a special case of Weil’s theorem as well.

Recall from section 2 that the solutions to $x^2 + y^2 = 1$, over \mathbb{R} , over \mathbb{Q} , or over various finite fields, could be parametrized by the variable m . It was this parametrization that enabled us to determine a simple formula for the size of $X(\mathbb{F}_p)$ in this case. We remarked earlier that most schemes could not be so parametrized; now we can make that statement a bit more precise, at least for algebraic curves.

Fact. If X is a genus-0 curve, then the points of X can be parametrized by a single variable.

The converse of this fact is more or less true as well (though stating it properly requires us to say more than we can here about “singular curves”). In other words, a

thoroughly algebraic question—whether the solutions of a Diophantine equation can be parametrized—is hereby given a geometric answer.

5.2 Rational Points on Curves

As we said above, some elliptic curves (which are curves of genus 1) have finitely many rational points, and others have infinitely many. What is the situation for algebraic curves of other flavors?

We have already encountered a curve of genus 0 with infinitely many points: namely, the curve $x^2 + y^2 = 1$. On the other hand, the curve $x^2 + y^2 = 7$ also has genus 0, and a simple modification of the argument of the first section shows that this curve has *no* rational points. It turns out these are the only two possibilities.

Fact. If X is a curve of genus 0, then $X(\mathbb{Q})$ is either empty or infinite.

Genus-1 curves are known to fall into a similar dichotomy, thanks to the theorem of Mazur we alluded to earlier.

Fact. If X is a genus-1 curve, then either X has at most sixteen rational points or it has infinitely many rational points.

What about curves of higher genus? In the early 1920s, Mordell made the following conjecture.

Conjecture. If X is a curve of genus greater than 2, then X has finitely many rational points.

This conjecture was proved by Faltings in 1983; in fact, he proved a more general theorem of which this conjecture is a special case. It is worth remarking that the work of Faltings involves a great deal of importation of geometric intuition to the study of the scheme $\text{Spec } \mathbb{Z}$.

When you prove that a set is finite, it is natural to wonder whether you can bound its size. For example, if $f(x)$ is a degree 6 polynomial with no repeated roots, the curve $y^2 = f(x)$ turns out to have genus 2; so by Faltings’s theorem there are only finitely many pairs of rational numbers (x, y) satisfying $y^2 = f(x)$.

Question. Is there a constant B such that, for all degree 6 polynomials with coefficients in \mathbb{Q} and no repeated roots, the equation $y^2 = f(x)$ has at most B solutions?

IV.6. Algebraic Topology

PUP: the proofreader wrote, 'The reader will get the context here?' To this non-mathematician, this reads like a non-sequitur.' Neither Tim nor I can see the problem - please explain.

This question remains open, and I do not think there is a strong consensus about whether the answer will be yes or no. The current world record is held by the curve $y^2 = 378\,371\,081x^2(x^2 - 9)^2 - 229\,833\,600(x^2 - 1)^2$, constructed by Keller and Kulesz, which has 588 rational points.

Interest in the above question comes from its relation to a conjecture of Lang, which involves points on higher-dimensional varieties. Caporaso, Harris, and Mazur showed that Lang's conjecture implies a positive answer to the question above. This suggests a natural attack on the conjecture: if one can find a way to construct an infinite sequence of degree 6 polynomials $f(x)$ so that the equations $y = f(x)$ have ever more numerous rational solutions, then one has a disproof of Lang's conjecture! No one has yet been successful at this task. If one could *prove* that the answer to the question above was affirmative, it would probably bolster our faith in the correctness of Lang's conjecture, though of course it would bring us no nearer to turning the conjecture into a theorem.

In this article we have seen only a glimpse of the modern theory of arithmetic geometry, and perhaps I have overemphasized mathematicians' successes at the expense of the much larger territory of questions, like Lang's conjecture above, about which we remain wholly ignorant. At this stage in the history of mathematics, we can confidently say that the schemes attached to Diophantine problems *have geometry*. What remains is to say as much as we can about *what this geometry is like*, and in this respect, despite the progress described here, our understanding is still quite unsatisfactory when compared with our knowledge of more classical geometric situations.

Further Reading

Dieudonné, J. 1985. *History of Algebraic Geometry*. Monterey, CA: Wadsworth.
Silverman, J., and J. Tate. 1992. *Rational Points on Elliptic Curves*. New York: Springer.

IV.6 Algebraic Topology

Burt Totaro

Introduction

Topology is concerned with the properties of a geometric shape that are unchanged when we continuously deform it. In more technical terms, topology tries to

classify TOPOLOGICAL SPACES [III.92], where two spaces are considered the same if they are homeomorphic. Algebraic topology assigns numbers to a topological space, which can be thought of as the "number of holes" in that space. These holes can be used to show that two spaces are not homeomorphic: if they have different numbers of holes of some kind, then one cannot be a continuous deformation of the other. In the happiest cases, we can hope to show the converse statement: that two spaces with the same number of holes (in some precise sense) *are* homeomorphic.

Topology is a relatively new branch of mathematics, with its origins in the nineteenth century. Before that, mathematics usually sought to solve problems exactly: to solve an equation, to find the path of a falling body, to compute the probability that a game of dice will lead to bankruptcy. As the complexity of mathematical problems grew, it became clear that most problems would never be solved by an exact formula: a classic example is the problem, known as THE THREE-BODY PROBLEM [V.36], of computing the future movements of Earth, the Sun, and the Moon under the influence of gravity. Topology allows the possibility of making qualitative predictions when quantitative ones are impossible. For example, a simple topological fact is that a trip from New York to Montevideo must cross the equator at some point, although we cannot say exactly where.

1 Connectedness and Intersection Numbers

Perhaps the simplest topological property is one called *connectedness*. This can be defined in various ways, as we shall see in a moment, but once we have a notion of what it means for a space to be connected we can then divide a topological space up into connected pieces, called *components*. The number of these pieces is a simple but useful INVARIANT [I.4 §2.2]: if two spaces have different numbers of connected components, then they are not homeomorphic.

For nice topological spaces, the different definitions of connectedness are equivalent. However, they can be generalized to give ways of measuring the number of holes in a space; these generalizations are interestingly different and all of them are important.

The first interpretation of connectedness uses the notion of a *path*, which is defined to be a continuous mapping f from the unit interval $[0, 1]$ to a given space X . (We think of f as a path from $f(0)$ to $f(1)$.) Let us declare two points of X to be equivalent if there is a path from one to the other. The set of EQUIVALENCE

CLASSES [I.2 §2.3] is called the set of *path components* of X and is written $\pi_0(X)$. This is a very natural way of defining the “number of connected pieces” into which X breaks up. One can generalize this notion by considering mappings into X from other standard spaces such as spheres: this leads to the notion of homotopy groups, which will be the topic of section 2.

A different way of thinking about connectedness is based on functions from X to the real line rather than functions from a line segment into X . Let us assume that we are in a situation where it makes sense to differentiate functions on X . For example, X could be an open subset of some Euclidean space, or more generally a SMOOTH MANIFOLD [I.3 §6.9]. Consider all the real-valued functions on X whose derivative is everywhere equal to zero: these functions form a real VECTOR SPACE [I.3 §2.3], which we call $H^0(X, \mathbb{R})$ (the “zeroth cohomology group of X with real coefficients”). Calculus tells us that if a function defined on an interval has derivative zero, then it must be constant, but that is not true when the domain has several connected pieces: all we can say then is that the function is constant on each connected piece of X . The number of degrees of freedom of such a function is therefore equal to the number of connected pieces, so the dimension of the vector space $H^0(X, \mathbb{R})$ is another way to describe the number of connected components of X . This is the simplest example of a cohomology group. Cohomology will be discussed in section 4.

We can use the idea of connectedness to prove a serious theorem of algebra: every real polynomial of odd degree has a real root. For example, there must be some real number x such that $x^3 + 3x - 4 = 0$. The basic observation is that when x is a large positive number or a highly negative number, the term x^3 is much bigger (in absolute value) than the other terms of the polynomial. Since this top term is an odd power of x , we have $f(x) > 0$ for some positive number x and $f(x) < 0$ for some negative number x . If f were never equal to zero, then it would be a continuous mapping from the real line into the real line minus the origin. But the real line is connected, while the real line minus the origin has two connected components, the positive and negative numbers. It is easy to show that a continuous map from a connected space X to another space Y must map X into just one connected component of Y : in our case, this contradicts the fact that f takes both positive and negative values. Therefore f must be equal to zero at some point, and the proof is complete.

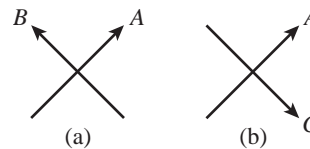


Figure 1 Intersection numbers:
(a) $A \cdot B = 1$; (b) $A \cdot C = -1$.

This argument can be phrased in terms of the “intermediate value theorem” of calculus, which is indeed one of the most basic topological theorems. An equivalent reformulation of this theorem states that a continuous curve that goes from the lower half-plane to the upper half-plane must cross the horizontal axis at some point. This idea leads to *intersection numbers*, one of the most useful concepts in topology. Let M be a smooth oriented manifold. (Roughly speaking, a manifold is oriented if you cannot continuously slide a shape about inside it and end up with a reflection of that shape. The simplest nonoriented manifold is a Möbius strip: to reflect a shape, slide it around the strip an odd number of times.) Let A and B be two closed oriented submanifolds of M with dimensions adding up to the dimension of M . Finally, suppose that A and B intersect transversely, so that their intersection has the “correct” dimension, namely 0, and is therefore a collection of separated points.

Now let p be one of these points. There is a way of assigning a weight of $+1$ or -1 to p , which depends in a natural way on the relationship between the orientations of A , B , and M (see figure 1). For example, if M is a sphere, A is the equator of M , B is a closed curve, and appropriate directions are given to A and B , then the weight of p will tell you whether B crosses A upwards or downwards at p . If A and B intersect in only finitely many points, then we can define the intersection number of A and B , written $A \cdot B$, to be the sum of the weights ($+1$ or -1) at all the intersection points. In particular, this will happen if M is COMPACT [III.9] (that is, we can think of it as a closed bounded subset of \mathbb{R}^N for some N).

The important point about the intersection number is that it is an *invariant*, in the following sense: if you move A and B about in a continuous way, ending up with another pair of transverse submanifolds A' and B' , then the intersection number $A' \cdot B'$ is the same as $A \cdot B$, even though the number of intersection points can change. To see why this might be true, consider again the case where A and B are curves and M is two

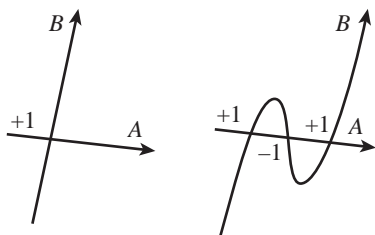


Figure 2 Moving a submanifold.

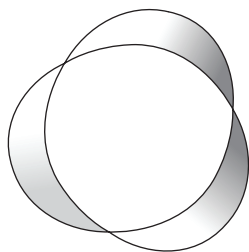


Figure 3 A surface bounded by a knot.

dimensional: if A and B meet at a point with weight 1, we can wiggle one of them to turn that point into three points with weights 1, -1 , and 1, but the total contribution to the intersection number is unchanged. This is illustrated in figure 2. As a result, the intersection number $A \cdot B$ is defined for *any* two submanifolds of complementary dimension: if they do not intersect transversely, one can move them until they do and use the definition we have just given.

In particular, if two submanifolds have nonzero intersection number, then they can never be moved to be disjoint from each other. This is another way to describe the earlier arguments about connectedness. It is easy to write down one curve from New York to Montevideo whose intersection number with the equator is equal to 1. Therefore, no matter how we move that curve (provided that we keep the endpoints fixed: more generally, if either A or B has a boundary, then that boundary should be kept fixed), its intersection number with the equator will always be 1, and in particular it must meet the equator in at least one point.

One of many applications of intersection numbers in topology is the idea of *linking numbers*, which comes from KNOT THEORY [III.46]. A *knot* is a path in space that begins and ends at the same point, or, more formally, a closed connected one-dimensional submanifold of \mathbb{R}^3 . Given any knot K , it is always possible to find a surface S in \mathbb{R}^3 with K as its boundary (see fig-

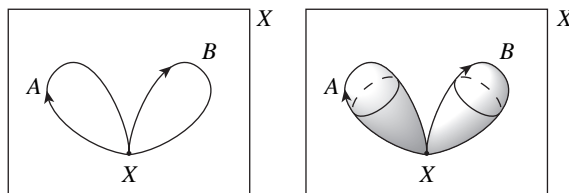


Figure 4 Multiplication in the fundamental group and in higher homotopy groups.

ure 3). Now let L be a knot that is disjoint from K . The linking number of K with L is defined to be the intersection number of L with the surface S . The properties of intersection numbers imply that if the linking number of K with L is nonzero, then the knots K and L are “linked,” in the sense that it is impossible to pull them apart.

2 Homotopy Groups

If we remove the origin from the plane \mathbb{R}^2 , then we obtain a new space that is different from the plane in a fundamental way: it has a hole in it. However, we cannot detect this difference by counting components, since both the plane and the plane without the origin are connected. We begin this section by defining an invariant called the *fundamental group*, which does detect this kind of hole.

As a first approximation, one could say that the elements of the fundamental group of a space X are *loops*, which can be formally defined as continuous functions f from $[0, 1]$ to X such that $f(0) = f(1)$. However, this is not quite accurate, for two reasons. The first reason, which is extremely important, is that two loops are regarded as equivalent if one can be continuously deformed to the other while all the time staying inside X . If this is the case, we say that they are *homotopic*. To be more formal about this, let us suppose that f_0 and f_1 are two loops. Then a *homotopy* between f_0 and f_1 is a collection of loops f_s in X , one for each s between 0 and 1, such that the function $F(s, t) = f_s(t)$ is a continuous function from $[0, 1]^2$ to X . Thus, as s increases from 0 to 1, the loop f_s moves continuously from f_0 to f_1 . If two loops are homotopic, then we count them as the same. So the elements of the homotopy group are not actually loops but equivalence classes, or *homotopy classes*, of loops.

Even this is not quite correct, because for technical reasons we need to impose an extra condition on our loops: that they all start from (and therefore end at)

a given point, called the *base point*. If X is connected, it turns out not to matter what this base point is, but we need it to be the same for all loops. The reason for this is that it gives us a way to multiply two loops: if x is the base point and A and B are two loops that start and end at x , then we can define a new loop by going around A and then going around B . This is illustrated in figure 4. We regard this new loop as the product of the loops A and B . It is not hard to check that the homotopy class of this product depends only on the homotopy classes of A and B , and that the resulting binary operation turns the set of homotopy classes of loops into a GROUP [I.3 §2.1]. It is this group that we call the fundamental group of X . It is denoted $\pi_1(X)$.

The fundamental group can be computed for most of the spaces we are likely to encounter. This makes it an important way to distinguish one space from another. First of all, for any n the fundamental group of \mathbb{R}^n is the trivial group with just one element, because any loop in \mathbb{R}^n can be continuously shrunk to its base point. On the other hand, the fundamental group of $\mathbb{R}^2 \setminus \{0\}$, the plane with the origin removed, is isomorphic to the group \mathbb{Z} of the integers. This tells us that we can associate with any loop in $\mathbb{R}^2 \setminus \{0\}$ an integer that does not change if we modify the loop in a continuous way. This integer is known as the *winding number*. Intuitively, the winding number measures the total number of times that the mapping goes around the origin, with counterclockwise circuits counting positively and clockwise ones negatively. Since the fundamental group of $\mathbb{R}^2 \setminus \{0\}$ is not the trivial group, $\mathbb{R}^2 \setminus \{0\}$ cannot be homeomorphic to the plane. (It is an interesting exercise to try to find an elementary proof of this result—that is, a proof that does not use, or implicitly reconstruct, any of the machinery of algebraic topology. Such proofs do exist, but it is tricky to find them.)

A classic application of the fundamental group is to prove THE FUNDAMENTAL THEOREM OF ALGEBRA [V.15], which states that every nonconstant polynomial with complex coefficients has a complex root. (The proof is sketched in the article just cited, though the fundamental group is not explicitly mentioned there.)

The fundamental group tells us about the number of “one-dimensional holes” that a space has. A basic example is given by the circle, which has fundamental group \mathbb{Z} , just as $\mathbb{R}^2 \setminus \{0\}$ does, and for essentially the same reason: given a path in the circle that begins and ends at the same point, we can see how many times it goes around the circle. In the next section we shall see some more examples.

Before we think about higher-dimensional holes, we first need to discuss one of the most important topological spaces: the n -dimensional sphere. For any natural number n , this is defined to be the set of points in \mathbb{R}^{n+1} at distance 1 from the origin. It is denoted S^n . Thus, the 0-sphere S^0 consists of two points, the 1-sphere S^1 is the circle, and the 2-sphere S^2 is the usual sphere, like the surface of Earth. Higher-dimensional spheres take a little bit of getting used to, but we can work with them in the same way that we can with lower-dimensional spheres. For example, we can construct the 2-sphere from a closed two-dimensional disk by identifying all the points on the boundary circle with each other. In the same way, the 3-sphere can be obtained from a solid three-dimensional ball by identifying all the points on the boundary 2-sphere. A related picture is to think of the 3-sphere as being obtained from our familiar three-dimensional space \mathbb{R}^3 by adding one point “at infinity.”

Now let us think about the familiar sphere S^2 . This has trivial fundamental group, since any loop drawn on the sphere can be shrunk to a point. However, this does not mean that the topology of S^2 is trivial. It just means that in order to detect its interesting properties we need a different invariant. And it is possible to base such an invariant on the observation that even if loops can always be shrunk, there are other maps that cannot. Indeed, the sphere itself cannot be shrunk to a point. To say this more formally, the identity map from the sphere to itself is not homotopic to a map from the sphere to just one point.

This idea leads to the notion of higher-dimensional homotopy groups of a topological space X . The rough idea is to measure the number of “ n -dimensional holes” in X , for any natural number n , by considering all the continuous mappings from the n -sphere to X . We want to see whether any of these spheres wrap around a hole in X . Once again, we consider two mappings from S^n to X to be equivalent if they are homotopic. And the elements of the n th homotopy group $\pi_n(X)$ are again defined to be the homotopy classes of these mappings.

Let f be a continuous map from $[0, 1]$ to X with $f(0) = f(1) = x$. If we like we can turn the interval $[0, 1]$ into the circle S^1 by “identifying” the points 0 and 1: then f becomes a map from S^1 to X , with one specified point in S^1 mapping to x . In order to be able to define a group operation for mappings from a higher-dimensional S^n , we similarly fix a point s in S^n and a base point x in X and look just at maps that send s to x .

Let A and B be two continuous mappings from S^n to X with this property. The “product” mapping $A \cdot B$ from S^n to X is defined as follows. First “pinch” the equator of S^n down to a point. When $n = 1$, the equator consists of just two points and the result is a figure eight. Similarly, for general n , we end up with two copies of S^n that touch each other, one made out of the northern hemisphere and one out of the southern hemisphere of the original unpinched copy of S^n . We now use the map A to map the bottom half into X and the map B to map the top half into X , with the equator mapping to the base point x . (For both halves, the pinched equator is playing the part of the point s .)

As in the one-dimensional case, this operation makes the set $\pi_n(X)$ into a group, and this group is the n th homotopy group of the space X . One can think of it as measuring how many “ n -dimensional holes” a space has.

These groups are the beginning of “algebraic” topology: starting from any topological space, we construct an algebraic object, in this case a group. If two spaces are homeomorphic, then their fundamental groups (and higher homotopy groups) must be isomorphic. This is richer than the original idea of just measuring the *number* of holes, since a group contains more information than just a number.

Any continuous function from S^n into \mathbb{R}^m can be continuously shrunk to a point in a straightforward way. This shows that all the higher homotopy groups of \mathbb{R}^m are also trivial, which is a precise formulation of the vague idea that \mathbb{R}^m has no holes.

Under certain circumstances one can show that two different topological spaces X and Y must have the same number of holes of all types. This is clearly true if X and Y are homeomorphic, but it is also true if X and Y are equivalent in a weaker sense, known as *homotopy equivalence*. Let X and Y be topological spaces and let f_0 and f_1 be continuous maps from X to Y . A homotopy from f_0 to f_1 is defined more or less as it was for spheres: it is a continuous family of continuous maps from X to Y that starts with f_0 and ends with f_1 . As then, if such a homotopy exists, we say that f_0 and f_1 are homotopic. Next, a homotopy equivalence from a space X to a space Y is a continuous map $f : X \rightarrow Y$ such that there is another continuous map $g : Y \rightarrow X$ with the property that the composition $g \circ f : X \rightarrow X$ is homotopic to the identity map on X , and $f \circ g : Y \rightarrow Y$ is homotopic to the identity map on Y . (Notice that if we replaced the word “homotopic” with “equal,” we would obtain the definition of a homeomorphism.) When there

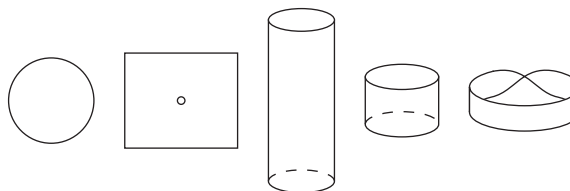


Figure 5 Some spaces that are homotopy equivalent to the circle.

is a homotopy equivalence from X to Y , we say that X and Y are *homotopy equivalent*, and also that X and Y have the same *homotopy type*.

A good example is when X is the unit circle and Y is the plane with the origin removed. We have already observed that these have the same fundamental group, and commented that it was “for essentially the same reason.” Now we can be more precise. Let $f : X \rightarrow Y$ be the map that takes (x, y) to (x, y) (where the first (x, y) belongs to the circle and the second to the plane). Let $g : Y \rightarrow X$ be the map that takes (u, v) to

$$\left(\frac{u}{\sqrt{u^2 + v^2}}, \frac{v}{\sqrt{u^2 + v^2}} \right).$$

(Note that $u^2 + v^2$ is never zero because the origin is not contained in Y .) Then $g \circ f$ is easily seen to equal the identity on the unit circle, so it is certainly homotopic to the identity. As for $f \circ g$, it is given by the same formula as g itself. More geometrically, it takes the points on each radial line to the point where that line intersects the unit circle. It is not hard to show that this map is homotopic to the identity on Y . (The basic idea is to “shrink the radial lines down” to the points where they intersect the circle.)

Very roughly speaking, two spaces are homotopy equivalent if they have the same number of holes of all types. This is a more flexible notion of “having the same shape” than the notion of homeomorphism. For example, Euclidean spaces of different dimensions are not homeomorphic to each other, but they are all homotopy equivalent. Indeed, they are all homotopy equivalent to a point: such spaces are called *contractible*, and one thinks of them as the spaces that have no hole of any sort. The circle is not contractible, but it is homotopy equivalent to many other natural spaces: the plane \mathbb{R}^2 minus the origin (as we have seen), the cylinder $S^1 \times \mathbb{R}$, the compact cylinder $S^1 \times [0, 1]$, and even the Möbius strip (see figure 5). Most invariants in algebraic topology (such as homotopy groups and cohomology groups) are the same for any two spaces that are homotopy equivalent. Thus, knowing that the fundamental

T&T note: PUP figure needs shading. Will speak to Dimitri about it in due course.

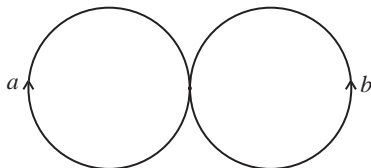


Figure 6 One-point union of two circles.

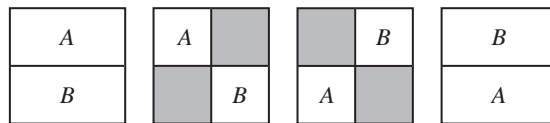
group of the circle is isomorphic to the integers tells us that the same is true for the various homotopy equivalent spaces just mentioned. Roughly speaking, this says that all these spaces have “one basic one-dimensional hole.”

3 Calculations of the Fundamental Group and Higher Homotopy Groups

To give some more feeling for the fundamental group, let us review what we already know and look at a few more examples. The fundamental group of the 2-sphere, or indeed of any higher-dimensional sphere, is trivial. The two-dimensional torus $S^1 \times S^1$ has fundamental group $\mathbb{Z}^2 = \mathbb{Z} \times \mathbb{Z}$. Thus, a loop in the torus determines two integers, which measure how many times it winds around in the meridian direction and how many in the longitudinal direction.

The fundamental group can also be non-Abelian; that is, we can have $ab \neq ba$ for some elements a and b of the fundamental group. The simplest example is a space X built out of two circles that meet at a single point (see figure 6). The fundamental group of X is the FREE GROUP [IV.10 §2] on two generators a and b . Roughly speaking, an element of this group is any product you can write down using the generators and their inverses, such as $abaab^{-1}a$, except that if a and a^{-1} or b and b^{-1} appear next to each other, you cancel them first. (So instead of $abb^{-1}bab^{-1}$ one would simply write $abab^{-1}$, for example.) The generators correspond to loops around each of the two circles. The free group is in a sense the most highly non-Abelian group. In particular, ab is not equal to ba , which in topological terms tells us that going around loop a and then loop b in the space X is not homotopic to the loop that goes around loop b and then loop a .

This space may seem somewhat artificial, but it is homotopy equivalent to the plane with two points removed, which appears in many contexts. More generally, the fundamental group of the plane with d points removed is the free group on d generators: this is a pre-

Figure 7 Proof that π_2 of any space is Abelian.

cise sense in which the fundamental group measures the number of holes.

In contrast with the fundamental group, the higher homotopy groups $\pi_n(X)$ are Abelian when n is at least 2. Figure 7 gives a “proof without words” in the case $n = 2$, the proof being the same for any larger n . In the figure, we view the 2-sphere as the square with its boundary identified to a point. So any elements A and B of $\pi_2(X)$ are represented by continuous maps of the square to X that map the boundary of the square to the base point x . The figure exhibits (several steps of) a homotopy from AB to BA , with the shaded regions and the boundary of the square all mapping to the base point x . The picture is reminiscent of the simplest nontrivial braid, in which one string is twisted around another; this is the beginning of a deep connection between algebraic topology and BRAID GROUPS [III.4].

The fundamental group is especially powerful in low dimensions. For example, every compact connected surface (or two-dimensional manifold) is homeomorphic to one of those on a standard list (see DIFFERENTIAL TOPOLOGY [IV.7 §2.3]), and we compute that all the manifolds on this list have different (nonisomorphic) fundamental groups. So, when you capture a closed surface in the wild, computing its fundamental group tells you exactly where it fits in the classification. Moreover, the geometric properties of the surface are closely tied to its fundamental group. The surfaces with a RIEMANNIAN METRIC [I.3 §6.10] of positive CURVATURE [III.13] (the 2-sphere and REAL PROJECTIVE PLANE [I.3 §6.7]) are exactly the surfaces with finite fundamental group; the surfaces with a metric of curvature zero (the torus and Klein bottle) are exactly the surfaces with a fundamental group that is infinite but “almost Abelian” (there is an Abelian subgroup of finite index); and the remaining surfaces, those that have a metric of negative curvature, have “highly non-Abelian” fundamental group, like the free group (see figure 8).

After more than a century of studying three-dimensional manifolds, we now know, thanks to the advances of Thurston and Perelman, that the picture is almost the same for these as it is for 2-manifolds: the fun-



Figure 8 A sphere, a torus, and a surface of genus 2.

damental group controls the geometric properties of the 3-manifold almost completely (see DIFFERENTIAL TOPOLOGY [IV.7 §2.4]). But this is completely untrue for 4-manifolds and in higher dimensions: there are many different *simply connected* manifolds, meaning manifolds with trivial fundamental group, and we need more invariants to be able to distinguish between them. (To begin with, the 4-sphere S^4 and the product $S^2 \times S^2$ are both simply connected. More generally, we can take the connected sum of any number of copies of $S^2 \times S^2$, obtained by removing 4-balls from these manifolds and identifying the boundary 3-spheres. These 4-manifolds are all simply connected, and yet no two of them are homeomorphic or even homotopy equivalent.)

An obvious approach to distinguishing different spaces would be to use *higher* homotopy groups, and indeed this works in simple cases. For example, π_2 of the connected sum of r copies of $S^2 \times S^2$ is isomorphic to \mathbb{Z}^{2r} . Also, we can show that the sphere S^n of any dimension is not contractible (although it is simply connected for $n \geq 2$) by computing that $\pi_n(S^n)$ is isomorphic to the integers (rather than the trivial group). Thus, each continuous map from the n -sphere to itself determines an integer, called the *degree* of the map, which generalizes the notion of winding number for maps from the circle to itself.

In general, however, the homotopy groups are not a practical way of distinguishing one space from another, because they are amazingly hard to compute. A first hint of this was Hopf's 1931 discovery that $\pi_3(S^2)$ is isomorphic to the integers: it is clear that the 2-sphere has a two-dimensional hole, as measured by $\pi_2(S^2) \cong \mathbb{Z}$, but in what sense does it have a three-dimensional hole? This does not correspond to our naive view of what such a hole should be. The problem of computing the homotopy groups of spheres turns out to be one of the hardest in all of mathematics: some of what we know is shown in table 1, but despite massive efforts the homotopy groups $\pi_i(S_2)$, for example, are known only for $i \leq 64$. There are tantalizing patterns in these calculations, with a number-theoretic flavor, but it seems impossible to formulate a precise

guess for the homotopy groups of spheres in general. And computing the homotopy groups for spaces more complex than spheres is even more complicated.

To get an idea of the difficulties involved, let us define the so-called *Hopf map* from S^3 to S^2 , which turns out to represent a nonzero element of $\pi_3(S^2)$. There are in fact several equivalent definitions. One of them is to regard a point (x_1, x_2, x_3, x_4) in S^3 as a pair of complex numbers (z_1, z_2) such that $|z_1|^2 + |z_2|^2 = 1$. This we do by setting $z_1 = x_1 + ix_2$ and $z_2 = x_3 + ix_4$. We then map the pair (z_1, z_2) to the complex number z_1/z_2 . This may not look like a map to S^2 , but it is because z_2 may be zero, so in fact the image of the map is not \mathbb{C} but the *Riemann sphere* $\mathbb{C} \cup \infty$, which can be identified with S^2 in a natural way.

Another way of defining the Hopf map is to regard points (x_1, x_2, x_3, x_4) in S^3 as unit quaternions. In the article on quaternions in this volume [III.78], it is shown that each unit quaternion can be associated with a rotation of the sphere. If we fix some point s in the sphere and map each unit quaternion to the image of s under the associated rotation, then we get a map from S^3 to S^2 that is homotopic to the map defined in the previous paragraph.

The Hopf map is an important construction, and will reappear more than once later in this article.

4 Homology Groups and the Cohomology Ring

Homotopy groups, then, can be rather mysterious and very hard to calculate. Fortunately, there is a different way to measure the number of holes in a topological space: homology and cohomology groups. The definitions are more subtle than the definition of homotopy groups, but the groups turn out to be easier to compute and are for this reason much more commonly used.

Recall that elements of the n th homotopy group $\pi_n(X)$ of a topological space X are represented by continuous maps from the n -sphere to X . Let X be a manifold, for simplicity. There are two key differences between homotopy groups and homology groups. The

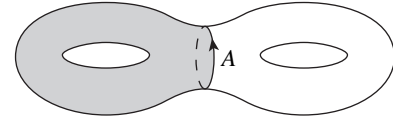
Table 1 The first few homotopy groups of spheres.

	S^1	S^2	S^3	S^4	S^5	S^6	S^7	S^8	S^9
π_1	\mathbb{Z}	0	0	0	0	0	0	0	0
π_2	0	\mathbb{Z}	0	0	0	0	0	0	0
π_3	0	\mathbb{Z}	\mathbb{Z}	0	0	0	0	0	0
π_4	0	$\mathbb{Z}/2$	$\mathbb{Z}/2$	\mathbb{Z}	0	0	0	0	0
π_5	0	$\mathbb{Z}/2$	$\mathbb{Z}/2$	$\mathbb{Z}/2$	\mathbb{Z}	0	0	0	0
π_6	0	$\mathbb{Z}/4 \times \mathbb{Z}/3$	$\mathbb{Z}/4 \times \mathbb{Z}/3$	$\mathbb{Z}/2$	$\mathbb{Z}/2$	\mathbb{Z}	0	0	0
π_7	0	$\mathbb{Z}/2$	$\mathbb{Z}/2$	$\mathbb{Z} \times \mathbb{Z}/4 \times \mathbb{Z}/3$	$\mathbb{Z}/2$	$\mathbb{Z}/2$	\mathbb{Z}	0	0
π_8	0	$\mathbb{Z}/2$	$\mathbb{Z}/2$	$\mathbb{Z}/2 \times \mathbb{Z}/2$	$\mathbb{Z}/8 \times \mathbb{Z}/3$	$\mathbb{Z}/2$	$\mathbb{Z}/2$	\mathbb{Z}	0
π_9	0	$\mathbb{Z}/3$	$\mathbb{Z}/3$	$\mathbb{Z}/2 \times \mathbb{Z}/2$	$\mathbb{Z}/2$	$\mathbb{Z}/8 \times \mathbb{Z}/3$	$\mathbb{Z}/2$	$\mathbb{Z}/2$	\mathbb{Z}
π_{10}	0	$\mathbb{Z}/3 \times \mathbb{Z}/5$	$\mathbb{Z}/3 \times \mathbb{Z}/5$	$\mathbb{Z}/8 \times \mathbb{Z}/3 \times \mathbb{Z}/3$	$\mathbb{Z}/2$	0	$\mathbb{Z}/8 \times \mathbb{Z}/3$	$\mathbb{Z}/2$	$\mathbb{Z}/2$

first is that the basic objects of homology are more general than n -dimensional spheres: *every* closed oriented n -dimensional submanifold A of X determines an element of the n th homology group of X , $H_n(X)$. This might make homology groups seem much bigger than homotopy groups, but that is not the case, because of the second major difference between homotopy and homology. As with homotopy, the elements of the homology groups are not the submanifolds themselves but equivalence classes of submanifolds, but the definition of the equivalence relation for homology makes it much easier for two of these submanifolds to be equivalent than it is for two spheres to be homotopic.

We shall not give a formal definition of homology, but here are some examples that convey some of its flavor. Let X be the plane with the origin removed and let A be a circle that goes around the origin. If we continuously deform this circle, we will obtain a new curve that is homotopic to the original circle, but with homology we can do more. For instance, we can start with a continuous deformation that causes two of its points to touch and turns it into a figure eight. One half of this figure eight will have to contain the origin, but we can leave that still and slide the other part away. The result is then two closed curves, with the origin inside one and outside the other. This pair of curves, which together form a 1-manifold with two components, is equivalent to the original circle. It can be seen as a continuous deformation of a more general kind.

A second example shows how natural it is to include other manifolds in the definition of homology. This time let X be \mathbb{R}^3 with a circle removed, and let A be a sphere that contains the circle in its interior. Suppose that the circle is in the XY -plane and that both it and the sphere A are centered at the origin. Then we can

**Figure 9** The circle A represents zero in the homology of the surface.

pinch the top and bottom of A toward the origin until they just touch. If we do so, then we obtain a shape that looks like a torus, except that the hole in the middle has been shrunk to zero. But we can open up this hole with the help of a further continuous deformation and obtain a genuine torus, which is a “tube” around the original circle. From the point of view of homology, this torus is equivalent to the sphere A .

A more general rule is that if X is a manifold and B is a compact oriented $(n + 1)$ -dimensional submanifold of X with a boundary, then this boundary ∂B will be equivalent to zero (which is the same as saying that $[\partial B] = 0$ in $H_n(X)$): see figure 9.

The group operation is easy to define: if A and B are two disjoint submanifolds of X , giving rise to homology classes $[A]$ and $[B]$, then $[A] + [B]$ is the homology class of $[A \cup B]$. (More generally, the definition of homology allows us to add up any collection of submanifolds, whether or not they overlap.) Here are some simple examples of homology groups, which, unlike the fundamental group, are always Abelian. The homology groups of a sphere, $H_i(S^n)$, are isomorphic to the integers \mathbb{Z} for $i = 0$ and for $i = n$, and 0 otherwise. This contrasts with the complicated homotopy groups of the sphere, and better reflects the naive idea that the n -sphere has one n -dimensional hole and no other holes. Note that the fundamental group of the circle, the group of integers, is the same as its first homology group. More generally, for any path-connected space,

the first homology group is always the “Abelianization” of the fundamental group (which is formally defined to be its largest Abelian quotient). For example, the fundamental group of the plane with two points removed is the free group on two generators, while the first homology group is the free *Abelian* group on two generators, or \mathbb{Z}^2 .

The homology groups of the two-dimensional torus $H_i(S^1 \times S^1)$ are isomorphic to \mathbb{Z} for $i = 0$, to \mathbb{Z}^2 for $i = 1$, and to \mathbb{Z} for $i = 2$. All of this has geometric meaning. The zeroth homology group of any space is isomorphic to \mathbb{Z}^r for a space X with r connected components. So the fact that the zeroth homology group of the torus is isomorphic to \mathbb{Z} means that the torus is connected. Any closed loop in the torus determines an element of the first homology group \mathbb{Z}^2 , which measures how many times the loop winds around the meridian and longitudinal directions of the torus. And finally, the homology of the torus in dimension 2 is isomorphic to \mathbb{Z} because the torus is a closed orientable manifold. That tells us that the whole torus defines an element of the second homology group of the torus, which is in fact a generator of that group. By contrast, the homotopy group $\pi_2(S^1 \times S^1)$ is the trivial group: there are no interesting maps from the 2-sphere to the 2-torus, but homology shows that there are interesting maps from other closed 2-manifolds to the 2-torus.

As we have mentioned, calculating homology groups is much easier than calculating homotopy groups. The main reason for this is the existence of results that tell you the homology groups of a space that is built up from smaller pieces in terms of the homology groups of those pieces and their intersections. Another important property of homology groups is that they are “functorial” in the sense that a continuous map f from a space X to a space Y leads in a natural way to a homomorphism f_* from $H_i(X)$ to $H_i(Y)$ for each i : $f_*([A])$ is defined to be $[f(A)]$. In other words, $f_*([A])$ is the equivalence class of the image of A under f .

We can define the closely related idea of “cohomology” simply by a different numbering. Let X be a closed oriented n -dimensional manifold. Then we define the i th *cohomology group* $H^i(X)$ to be the homology group $H_{n-i}(X)$. Thus, one way to write down a cohomology class (an element of $H^i(X)$) is by choosing a closed oriented submanifold S of codimension i in X . (This means that the dimension of S is $n - i$.) We write $[S]$ for the corresponding cohomology class.

For more general spaces than manifolds, cohomology is not just a simple renumbering of homology. Infor-

mally, if X is a topological space, then we think of an element of $H^i(X)$ as being represented by a codimension- i subspace of X that can move around freely in X . For example, suppose that f is a continuous map from X to an i -dimensional manifold. If X is a manifold and f is sufficiently “well-behaved,” then the inverse image of a “typical” point in the manifold will be an i -codimensional submanifold of X , and as we move the point about, this submanifold will vary continuously, and will do so in a way that is similar to the way that a circle became two circles and a sphere became a torus earlier. If X is a more general topological space, the map f still determines a cohomology class in $H^i(X)$, which we think of as being represented by the inverse image in X of any point in the manifold.

However, even when X is an oriented n -dimensional manifold, cohomology has distinct advantages over homology. This may seem odd, since the cohomology groups are the homology groups with different names. However, this renumbering allows us to give very useful extra algebraic structure to the cohomology groups of X : not only can we add cohomology classes, we can multiply them as well. Furthermore, we can do so in such a way that, taken together, the cohomology groups of X form a RING [III.83 §1]. (Of course, we could do this for the homology groups, but the cohomology groups form a so-called *graded* ring. In particular, if $[A] \in H^i(X)$ and $[B] \in H^j(X)$, then $[A] \cdot [B] \in H^{i+j}(X)$.)

The multiplication of cohomology classes has a rich geometric meaning, especially on manifolds: it is given by the *intersection* of two submanifolds. This generalizes our discussion of intersection numbers in section 1: there we considered zero-dimensional intersections of submanifolds, whereas we are now considering (cohomology classes of) higher-dimensional intersections. To be precise, let S and T be closed oriented submanifolds of X , of codimension i and j , respectively. By moving S slightly (which does not change its class in $H^i(X)$) we can assume that S and T intersect transversely, which implies that the intersection of S and T is a smooth submanifold of codimension $i + j$ in X . Then the product of the cohomology classes $[S]$ and $[T]$ is simply the cohomology class of the intersection $S \cap T$ in $H^{i+j}(X)$. (In addition, the submanifold $S \cap T$ inherits an orientation from S , T , and X : this is needed to define the associated cohomology class.)

As a result, to compute the cohomology ring of a manifold, it is enough to specify a basis for the cohomology groups (which, as we have already discussed, are relatively easy to determine) using some submanifolds

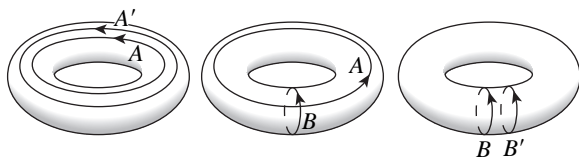


Figure 10 $A^2 = A \cdot A' = 0$, $A \cdot B = [\text{point}]$,
and $B^2 = B \cdot B' = 0$.

and to see how these submanifolds intersect. For example, we can compute the cohomology ring of the 2-torus as shown in figure 10. For another example, it is not hard to show that the cohomology of the COMPLEX PROJECTIVE PLANE [III.74] \mathbb{CP}^2 has a basis given by three basic submanifolds: a point, which belongs to $H^4(\mathbb{CP}^2)$ because it is a submanifold of codimension 4; a complex projective line $\mathbb{CP}^1 = S^2$, which belongs to $H^2(\mathbb{CP}^2)$; and the whole manifold \mathbb{CP}^2 , which is in $H^0(\mathbb{CP}^2)$ and represents the identity element 1 of the cohomology ring. The product in the cohomology ring is described by saying that $[\mathbb{CP}^1][\mathbb{CP}^1] = [\text{point}]$, because any two distinct lines \mathbb{CP}^1 in the plane meet transversely in a single point.

This calculation of the cohomology ring of the complex projective plane, although very simple, has several strong consequences. First of all, it implies Bézout's theorem on intersections of complex algebraic curves (see ALGEBRAIC GEOMETRY [IV.4 §6]). An algebraic curve of degree d in \mathbb{CP}^2 represents d times the class of a line \mathbb{CP}^1 in $H^2(\mathbb{CP}^2)$. Therefore, if two algebraic curves D and E of degrees d and e meet transversely, then the cohomology class $[D \cap E]$ equals

$$[D] \cdot [E] = (d[\mathbb{CP}^1])(e[\mathbb{CP}^1]) = de[\text{point}].$$

For complex submanifolds of a complex manifold, intersection numbers are always +1, not -1, and so this means that D and E meet in exactly de points.

We can also use the computation of the cohomology ring of \mathbb{CP}^2 to prove something about the homotopy groups of spheres. It turns out that \mathbb{CP}^2 can be constructed as the union of the 2-sphere and the closed four-dimensional ball, with each point of the boundary S^3 of the ball identified with a point in S^2 by the Hopf map, which was defined in the previous section.

A constant map from one space to another, or a map homotopic to a constant map, gives rise to the zero homomorphism between the homology groups H_i , at least when $i > 0$. The Hopf map $f : S^3 \rightarrow S^2$ also induces the zero homomorphism because the nonzero homology groups of S^3 and S^2 are in different dimen-

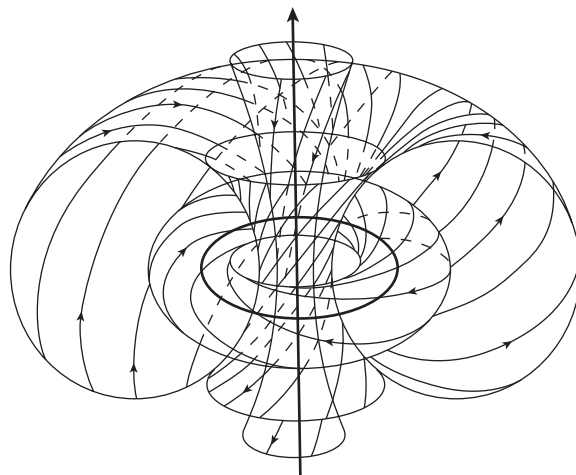


Figure 11 Fibers of the Hopf map.

sions. Nonetheless, we will show that f is not homotopic to the constant map. If it were, then the space \mathbb{CP}^2 obtained by attaching a 4-ball to the 2-sphere using the map f would be homotopy equivalent to the space obtained by attaching a 4-ball to the 2-sphere using a constant map. The latter space Y is the union of S^2 and S^4 identified at one point. But in fact Y is not homotopy equivalent to the complex projective plane, because their cohomology rings are not isomorphic. In particular, the product of any element of $H^2(Y)$ with itself is zero, unlike what happens in \mathbb{CP}^2 where $[\mathbb{CP}^1][\mathbb{CP}^1] = [\text{point}]$. Therefore f is nonzero in $\pi_3(S^2)$. A more careful version of this argument shows that $\pi_3(S^2)$ is isomorphic to the integers, and the Hopf map $f : S^3 \rightarrow S^2$ is a generator of this group.

This argument shows some of the rich relations between all the basic concepts of algebraic topology: homotopy groups, cohomology rings, manifolds, and so on. To conclude, here is a way to visualize the non-triviality of the Hopf map $f : S^3 \rightarrow S^2$. Look at the subset of S^3 that maps to any given point of the 2-sphere. These inverse images are all circles in the 3-sphere. To draw them, we can use the fact that S^3 minus a point is homeomorphic to \mathbb{R}^3 ; so these inverse images form a family of disjoint circles that fills up three-dimensional space, with one circle being drawn as a line (the circle through the point we removed from S^3). The striking feature of this picture is that any two of this huge family of circles have linking number 1 with each other: there is no way to pull any two of them apart (see figure 11).

PUP: Tim thinks that the use of 'fibers' in the caption is not likely to confuse the reader.

5 Vector Bundles and Characteristic Classes

We now introduce another major topological idea: fiber bundles. If E and B are topological spaces, x is a point in B , and $p : E \rightarrow B$ is a continuous map, then the *fiber* of p over x is the subspace of E that maps to x . We say that p is a *fiber bundle*, with fiber F , if every fiber of p is homeomorphic to the same space F . We call B the *base space* and E the *total space*. For example, any product space $B \times F$ is a fiber bundle over B , called the trivial F -bundle over B . (The continuous map in this case is the map that takes (x, y) to x .) But there are many nontrivial fiber bundles. For example, the Möbius strip is a fiber bundle over the circle with fiber a closed interval. This example helps to explain the old name “twisted product” for fiber bundles. Another example: the Hopf map makes the 3-sphere the total space of a circle bundle over the 2-sphere.

Fiber bundles are a fundamental way to build up complicated spaces from simple pieces. We will focus on the most important special case: vector bundles. A *vector bundle* over a space B is a fiber bundle $p : E \rightarrow B$ whose fibers are all real vector spaces of some dimension n . This dimension is called the *rank* of the vector bundle. A *line bundle* means a vector bundle of rank 1; for example, we can view the Möbius strip (not including its boundary) as a line bundle over the circle S^1 . It is a *nontrivial* line bundle; that is, it is not isomorphic to the trivial line bundle $S^1 \times \mathbb{R}$. (There are many ways of constructing it: one is to take the strip $\{(x, y) : 0 \leq x \leq 1\}$ and identify each point $(0, y)$ with the point $(1, -y)$. The base space of this line bundle is the set of all points $(x, 0)$, which is a circle since $(0, 0)$ and $(1, 0)$ have been identified.)

If M is a smooth manifold of dimension n , its *tangent bundle* $TM \rightarrow M$ is a vector bundle of rank n . We can easily define this bundle by considering M as a submanifold of some Euclidean space \mathbb{R}^N . (Every smooth manifold can be embedded into Euclidean space.) Then TM is the subspace of $M \times \mathbb{R}^N$ of pairs (x, v) such that the vector v is tangent to M at the point x ; the map $TM \rightarrow M$ sends a pair (x, v) to the point x . The fiber over x then has the form of the set of all pairs (x, v) with v belonging to an affine subspace of \mathbb{R}^N of dimension equal to that of M . For any fiber bundle, a *section* means a continuous map from the base space B to the total space E that maps each point x in B to some point in the fiber over x . A section of the tangent bundle of a manifold is called a *vector field*. We can draw a vector

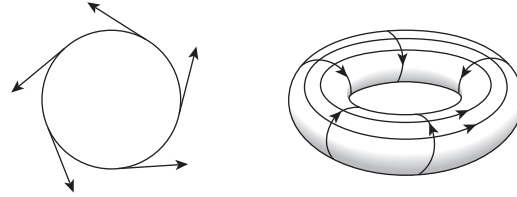


Figure 12 Trivializations of the tangent bundle for the circle and the torus.

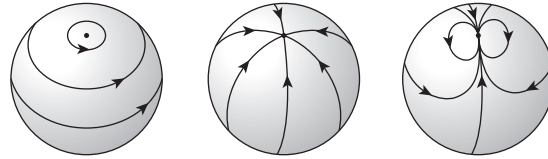


Figure 13 The hairy ball theorem.

field on a given manifold by putting an arrow (possibly of zero length) at every point of the manifold.

In order to classify smooth manifolds, it is important to study their tangent bundles, and in particular to see whether they are trivial or not. Some manifolds, like the circle S^1 and the torus $S^1 \times S^1$, do have trivial tangent bundle. The tangent bundle of an n -manifold M is trivial if and only if we can find n vector fields that are linearly independent at every point of M . So we can prove that the tangent bundle is trivial just by writing down such vector fields; see figure 12 for the circle or the torus. But how can we show that the tangent bundle of a given manifold is nontrivial?

One way is to use intersection numbers. Let M be a closed oriented n -manifold. We can identify M with the image of the “zero-section” inside the tangent bundle TM , the section that assigns to every point of M the zero vector at that point. Since the dimension of TM is precisely double that of M , the discussion of intersection numbers in section 1 gives a well-defined integer $M^2 = M \cdot M$, the self-intersection number of M inside TM ; this is called the *Euler characteristic* $\chi(M)$. By the definition of intersection numbers, for any vector field v on M that meets the zero-section transversely, the Euler characteristic of M is equal to the number of zeros of v , counted with signs.

As a result, if the Euler characteristic of M is not zero, then every vector field on M must meet the zero-section; in other words, every vector field on M must equal zero somewhere. The simplest example occurs when M is the 2-sphere S^2 . We can easily write down a vector field (for example, the one pointing toward

the east along circles of latitude, which vanishes at the north and south poles) whose intersection number with the zero-section is 2. Therefore the 2-sphere has Euler characteristic 2, and so every vector field on the 2-sphere must vanish somewhere. This is a famous theorem of topology known as the “hairy ball theorem”: it is impossible to comb the hair on a coconut (see figure 13).

This is the beginning of the theory of *characteristic classes*, which measure how nontrivial a given vector bundle is. There is no need to restrict ourselves to the tangent bundle of a manifold. For any oriented vector bundle E of rank n on a topological space X , we can define a cohomology class $\chi(E)$ in $H^n(X)$, the *Euler class*, which vanishes if the bundle is trivial. Intuitively, the Euler class of E is the cohomology class represented by the zero set of a general section of E , which (for example, if X is a manifold) should be a codimension- n submanifold of X , since X has codimension n in E . If X is a closed oriented n -manifold, then the Euler class of the tangent bundle in $H^n(X) = \mathbb{Z}$ is the Euler characteristic of X .

One inspiration for the theory of characteristic classes was the Gauss-Bonnet theorem, generalized to all dimensions in the 1940s. The theorem expresses the Euler characteristic of a closed manifold with a Riemannian metric as the integral over the manifold of a certain curvature function. More broadly, a central goal of differential geometry is to understand how the geometric properties of a Riemannian manifold such as its curvature are related to the topology of the manifold.

The characteristic classes for *complex* vector bundles (that is, bundles where the fibers are complex vector spaces) turn out to be particularly convenient: indeed, real vector bundles are often studied by constructing the associated complex vector bundle. If E is a complex vector bundle of rank n over a topological space X , the *Chern classes* of E are a sequence $c_1(E), \dots, c_n(E)$ of cohomology classes on X , with $c_i(E)$ belonging to $H^{2i}(X)$, which all vanish if the bundle is trivial. The top Chern class, $c_n(E)$, is simply the Euler class of E : thus, it is the first obstruction to finding a section of E that is everywhere nonzero. The more general Chern classes have a similar interpretation. For any $1 \leq j \leq n$, choose j general sections of E . The subset of X over which these sections become linearly dependent will have codimension $2(n+1-j)$ (assuming, for example, that X is a manifold). The Chern class $c_{n+1-j}(E)$ is precisely the cohomology class of this subset. Thus the Chern classes measure in a natural way the failure of a

given complex vector bundle to be trivial. The *Pontryagin classes* of a real vector bundle are defined to be the Chern classes of the associated complex vector bundle.

A triumph of differential topology is Sullivan’s 1977 theorem that there are only finitely many smooth closed simply connected manifolds of dimension at least 5 with any given homotopy type and given Pontryagin classes of the tangent bundle. This statement fails badly in dimension 4, as Donaldson discovered in the 1980s (see DIFFERENTIAL TOPOLOGY [IV.7 §2.5]).

6 K-Theory and Generalized Cohomology Theories

The effectiveness of vector bundles in geometry led to a new way of measuring the “holes” in a topological space X : looking at how many different vector bundles over X there are. This idea gives a simple way to define a cohomology-like ring associated to any space, known as *K-theory* (after the German word “Klasse,” since the theory involves equivalence classes of vector bundles). It turns out that *K-theory* gives a very useful new angle by which to look at topological spaces. Some problems that could be solved only with enormous effort using ordinary cohomology became easy with *K-theory*. The idea was created in algebraic geometry by Grothendieck in the 1950s and then brought into topology by Atiyah and Hirzebruch in the 1960s.

The definition of *K-theory* can be given in a few lines. For a topological space X , we define an Abelian group $K^0(X)$, the *K-theory* of X , whose elements can be written as formal differences $[E] - [F]$, where E and F are any two complex vector bundles over X . The only relations we impose in this group are that $[E \oplus F] = [E] + [F]$ for any two vector bundles E and F over X . Here $E \oplus F$ denotes the *direct sum* of the two bundles; if E_x and F_x denote the fibers at a given point x in X , the fiber of $E \oplus F$ at x is simply $E_x \times F_x$.

This simple definition leads to a rich theory. First of all, the Abelian group $K^0(X)$ is in fact a ring: we multiply two vector bundles on X by forming the *TENSOR PRODUCT* [III.91]. In this respect, *K-theory* behaves like ordinary cohomology. The analogy suggests that the group $K^0(X)$ should form part of a whole sequence of Abelian groups $K^i(X)$, for integers i , and indeed these groups can be defined. In particular, $K^{-i}(X)$ can be defined as the subgroup of those elements of $K^0(S^i \times X)$ whose restriction to $K^0(\text{point} \times X)$ is zero.

Then a miracle occurs: the groups $K^i(X)$ turn out to be *periodic* of order 2: $K^i(X) = K^{i+2}(X)$ for all integers

i. This is a famous phenomenon known as *Bott periodicity*. So there are really only two different K -groups attached to any topological space: $K^0(X)$ and $K^1(X)$.

This may suggest that K -theory contains less information than ordinary cohomology, but that is not so. Neither K -theory nor ordinary cohomology determines the other, although there are strong relations between them. Each brings different aspects of the shape of a space to the fore. Ordinary cohomology, with its numbering, shows fairly directly the way a space is built up from pieces of different dimensions. K -theory, having only two different groups, looks cruder at first (and is often easier to compute as a result). But geometric problems involving vector bundles often involve information that is subtle and hard to extract from ordinary cohomology, whereas this information is brought to the surface by K -theory.

The basic relation between K -theory and ordinary cohomology is that the group $K^0(X)$ constructed from the vector bundles on X “knows” something about all the even-dimensional cohomology groups of X . To be precise, the rank of the Abelian group $K^0(X)$ is the sum of the ranks of all the even-dimensional cohomology groups $H^{2i}(X)$. This connection comes from associating with a given vector bundle on X its Chern classes. The odd K -group $K^1(X)$ is related in the same way to the odd-dimensional ordinary cohomology.

As we have already hinted, the precise group $K^0(X)$, as opposed to just its rank, is better adapted to some geometric problems than ordinary cohomology. This phenomenon shows the power of looking at geometric problems in terms of vector bundles, and thus ultimately in terms of linear algebra. Among the classic applications of K -theory is the proof, by Bott, Kervaire, and Milnor, that the 0-sphere, the 1-sphere, the 3-sphere, and the 7-sphere are the only spheres whose tangent bundles are trivial. This has a deep algebraic consequence, in the spirit of the fundamental theorem of algebra: the only dimensions in which there can be a real division algebra (not assumed to be commutative or even associative) are 1, 2, 4, and 8. There are indeed division algebras of all four types: the real numbers, complex numbers, quaternions, and octonions (see QUATERNIONS, OCTONIONS, AND NORMED DIVISION ALGEBRAS [III.78]).

Let us see why the existence of a real division algebra of dimension n implies that the $(n - 1)$ -sphere has trivial tangent bundle. In fact, let us merely assume that we have a finite-dimensional real vector space V with a bilinear map $V \times V \rightarrow V$, which we call the “product,”

such that if x and y are vectors in V with $xy = 0$, then either $x = 0$ or $y = 0$. For convenience, let us also assume that there is an identity element 1 in V , so $1 \cdot x = x \cdot 1 = x$ for all $x \in V$; one can, however, do without this assumption. If V has dimension n , then we can identify V with \mathbb{R}^n . Then, for each point x in the sphere S^{n-1} , left multiplication by x gives a linear isomorphism from \mathbb{R}^n to itself. By scaling the output to have length 1, left multiplication by x gives a diffeomorphism from S^{n-1} to itself which maps the point 1 (scaled to have length 1) to x . Taking the derivative of this diffeomorphism at the point 1 gives a linear isomorphism from the tangent space of the sphere at the point 1 to the tangent space at x . Since the point x on the sphere is arbitrary, a choice of basis for the tangent space of the sphere at the point 1 determines a trivialization of the whole tangent bundle of the $(n - 1)$ -sphere.

Among other applications, K -theory provides the best “explanation” for the low-dimensional homotopy groups of spheres, and in particular for the number-theoretic patterns that are seen there. Notably, denominators of Bernoulli numbers appear among those groups (such as $\pi_{n+3}(S^n) \cong \mathbb{Z}/24$ for n at least 5), and this pattern was explained using K -theory by Milnor, Kervaire, and Adams.

THE ATIYAH-SINGER INDEX THEOREM [V.2] provides a deep analysis of linear differential equations on closed manifolds using K -theory. The theorem has made K -theory important for gauge theories and string theories in physics. K -theory can also be defined for noncommutative rings, and is in fact the central concept in “noncommutative geometry” (see OPERATOR ALGEBRAS [IV.15 §5]).

The success of K -theory led to a search for other “generalized cohomology theories.” There is one other theory that stands out for its power: *complex cobordism*. The definition is very geometric: the complex cobordism groups of a manifold M are generated by mappings of manifolds (with a complex structure on the tangent bundle) into M . The relations say that any manifold counts as zero if it is the boundary of some other manifold. For example, the union of two circles would count as zero if you could find a cylinder whose ends were those circles.

It turns out that complex cobordism is much richer than either K -theory or ordinary cohomology. It sees far into the structure of a topological space, but at the cost of being difficult to compute. Over the past thirty years, a whole series of cohomology theories,

such as elliptic cohomology and Morava K -theories, have been constructed as “simplifications” of complex cobordism: there is a constant tension in topology between invariants that carry a lot of information and invariants that are easy to compute. In one direction, complex cobordism and its variants provide the most powerful tool for the computation and understanding of the homotopy groups of spheres. Beyond the range where Bernoulli numbers appear, we see deeper number theory such as MODULAR FORMS [III.61]. In another direction, the geometric definition of complex cobordism makes it useful in algebraic geometry.

7 Conclusion

The line of thought introduced by pioneering topologists like RIEMANN [VI.49] is simple but powerful. Try to translate any problem, even a purely algebraic one, into geometric terms. Then ignore the details of the geometry and study the underlying shape or topology of the problem. Finally, go back to the original problem and see how much has been gained. The fundamental topological ideas such as cohomology are used throughout mathematics, from number theory to string theory.

Further Reading

From the definition of topological spaces to the fundamental group and a little beyond, I like M. A. Armstrong’s *Basic Topology* (Springer, New York, 1983). The current standard graduate textbook is A. Hatcher’s *Algebraic Topology* (Cambridge University Press, Cambridge, 2002). Two of the great topologists, Bott and Milnor, are also brilliant writers. Every young topologist should read R. Bott and L. Tu’s *Differential Forms in Algebraic Topology* (Springer, New York, 1982), J. Milnor’s *Morse Theory* (Princeton University Press, Princeton, NJ, 1963), and J. Milnor and J. Stasheff’s *Characteristic Classes* (Princeton University Press, Princeton, NJ, 1974).

IV.7 Differential Topology

C. H. Taubes

1 Smooth Manifolds

This article is about classifying certain objects called smooth manifolds, so I need to start by telling you what they are. A good example to keep in mind is the surface of a smooth ball. If you look at a small portion of it from very close up, then it looks like a portion of a

flat plane, but of course it differs in a radical way from a flat plane on larger distance scales. This is a general phenomenon: a smooth manifold can be very convoluted, but must be quite regular in close-up. This “local regularity” is the condition that each point in a manifold belongs to a neighborhood that looks like a portion of standard Euclidean space in some dimension. If the dimension in question is d for every point of the manifold, then the manifold itself is said to have dimension d . A schematic of this is shown in figure 1.

What does it mean to say that a neighborhood “looks like a portion of standard Euclidean space”? It means that there is a “nice” one-to-one map ϕ from the neighborhood into \mathbb{R}^d (with its usual notion of distance (see METRIC SPACES [III.58])). One can think of ϕ as “identifying” points in the neighborhood with points in \mathbb{R}^d : that is, x is identified with $\phi(x)$. If we do this, then the function ϕ is called a *coordinate chart* of the neighborhood, and any chosen basis for the linear functions on the Euclidean space is called a *coordinate system*. The reason for this is that ϕ allows us to use the coordinates in \mathbb{R}^d to label points in the neighborhood: if x belongs to the neighborhood, then one can label it with the coordinates of $\phi(x)$. For example, Europe is part of the surface of a sphere. A typical map of Europe identifies each point in Europe with a point in flat, two-dimensional Euclidean space, that is, a square grid labeled with latitude and longitude. These two numbers give us a coordinate system for the map, which can also be transferred to a coordinate system for Europe itself.

Now, here is a straightforward but central observation. Suppose that M and N are two neighborhoods that intersect, and suppose that functions $\phi : M \rightarrow \mathbb{R}^d$ and $\psi : N \rightarrow \mathbb{R}^d$ are used to give them each a coordinate chart. Then the intersection $M \cap N$ is given *two* coordinate charts, and this gives us an identification between the open regions $\phi(M \cap N)$ and $\psi(M \cap N)$ of \mathbb{R}^d : given a point x in the first region, the corresponding point in the second is $\psi(\phi^{-1}(x))$. This composition of maps is called a *transition function*, and it tells you how the coordinates from one of the charts on the intersecting region relate to those of the other. The transition function is a HOMEOMORPHISM [III.92] between the regions $\phi(M \cap N)$ and $\psi(M \cap N)$.

Suppose that you take a rectangular grid in the first Euclidean region and use the transition function $\psi\phi^{-1}$ to map it to the second one. It is possible that the image will again be a rectangular grid, but in general it will be somewhat distorted. An illustration is given in figure 2.

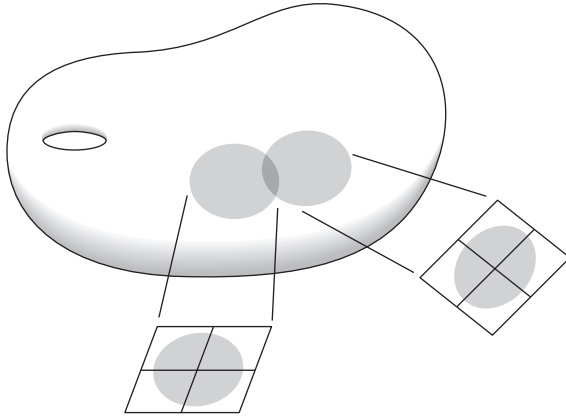


Figure 1 Small portions of a manifold resemble regions in a Euclidean space.

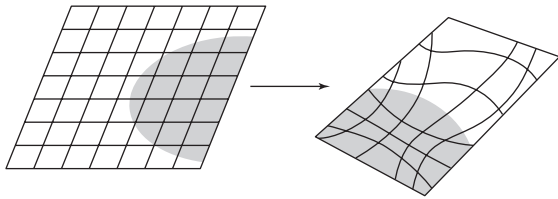


Figure 2 A transition function from a rectangular grid to a distorted rectangular grid.

The proper term for a space whose points are surrounded by regions that can be identified with parts of Euclidean space is a *topological manifold*. The word “topological” is used in order to indicate that there are no constraints on the coordinate-chart transition functions apart from the basic one that they should be continuous. However, some continuous functions are quite unpleasant, so one typically introduces extra constraints in order to limit the distorting effect that the transition functions can have on a rectangular coordinate grid.

Of prime interest here is the case where the transition functions are required to be differentiable to all orders. If a manifold has a collection of charts for which all the transition functions are infinitely differentiable, then it is said to have a *smooth structure*, and it is called a *smooth manifold*. Smooth manifolds are especially interesting because they are the natural arena for calculus. Roughly speaking, they are the most general context in which the notion of differentiation to any order makes intrinsic sense.

A function f , defined on a manifold, is said to be *differentiable* if, given any of its coordinate charts $\phi : N \rightarrow \mathbb{R}^d$, the function $g(y) = f(\phi^{-1}(y))$ (which is defined on a region of \mathbb{R}^d) is DIFFERENTIABLE [I.3 §5.3]. Calculus is impossible on a manifold if it does not admit charts with differentiable transition functions, because a function that might appear differentiable in one chart will not, in general, be differentiable when viewed from a neighboring chart.

Here is a one-dimensional example to illustrate this point. Consider the following two coordinate charts of a neighborhood of the origin in the real line. The first is the obvious chart that simply represents a real number x by itself. (Formally speaking, one is taking the function ϕ to be defined by the simple formula $\phi(x) = x$.) The second represents x by the point $x^{1/3}$. (Here the cube root of a negative number x is defined to be minus the cube root of $-x$.) What is the transition function between these two charts? Well, if t is a point in the region of \mathbb{R} used for the first chart, then $\phi^{-1}(t) = t$, so $\psi(\phi^{-1}(t)) = \psi(t) = t^{1/3}$. This is a continuous function of t but it is not differentiable at the origin.

Now consider the simplest possible function defined on the region used for the second chart, the function $h(s) = s$, and let us work out the corresponding function f on the manifold itself. The value of f at x should be the value of h at the point s corresponding to x . This point is $\psi(x) = x^{1/3}$, so $f(x) = h(x^{1/3}) = x^{1/3}$. Finally, since the point x in the manifold corresponds to the point $t = \phi(x) = x$ in the first region, the corresponding function on the first region is $g(t) = t^{1/3}$. (This is the same function as f only because ϕ happens to be the very special map that takes each number to itself.) Thus, the eminently differentiable function h on one coordinate chart translates into the continuous but not differentiable function g on the other.

Suppose one is given a topological manifold M with two sets of charts, both of which have infinitely differentiable transition functions. Then each set of charts gives us a smooth structure on the manifold. Of great importance is the fact that these two smooth structures can be fundamentally different.

To see what this means, let us call the sets of charts K and L . Given a function f , let us call it *K-differentiable* if it is differentiable from the viewpoint of K , and *L-differentiable* if it is differentiable from the viewpoint of L . It may easily happen that a function is K -differentiable without being L -differentiable or vice versa. However, we can say that K and L give the same smooth structure on M when there is a map, F ,

from M to itself with the following three properties. First, F is invertible and both F and F^{-1} are continuous. Second, the composition of F with any function that is K -differentiable is L -differentiable. Third, the composition of the inverse of F^{-1} with any function that is L -differentiable is K -differentiable. Loosely speaking, F turns the K -differentiable functions into L -differentiable ones and F^{-1} turns them back again. If no such function F exists, then the smooth structures given by K and L are considered to be genuinely different.

To see how this plays out, let us look at the one-dimensional example again. As noted previously, the functions that you deem to be differentiable if you use the ϕ -chart are not the same as those you deem to be differentiable if you use the ψ -chart. For example, the function $x \mapsto x^{1/3}$ is not ϕ -differentiable but it is ψ -differentiable. Even so, the ϕ -differentiable and ψ -differentiable sets of functions define the *same* smooth structure for the line, since any ψ -differentiable function becomes ϕ -differentiable once you compose it with the self-map $F : t \mapsto t^3$.

It is very far from obvious that any manifold can have more than one smooth structure, but this turns out to be the case. There are also manifolds that are entirely lacking in smooth structures. These two facts lead directly to the central concern of this essay, the long-sought quest for the two holy grails of differential topology.

- A list of all smooth structures on any given topological manifold.
- An algorithm to identify any given smooth structure on any given topological manifold with the corresponding structure from the list.

2 What Is Known about Manifolds?

Much has been accomplished as of the writing of this article with respect to the two points listed above. This said, the task for this part of the article is to summarize the state of affairs at the beginning of the twenty-first century. Various examples of manifolds are described along the way.

The story here requires a brief, preliminary digression to set the stage. If you have two manifolds and you set them side by side without their touching, then technically speaking they can be regarded as a single manifold that happens to have two components. In such a case, one can study the components individually. Therefore, in this article I shall talk exclusively about

connected manifolds: that is, manifolds with just one component. In a connected manifold, one can get from any point to any other point without ever leaving the manifold.

A second technical point is that it is useful to distinguish between manifolds such as the sphere, which are bounded in extent, and manifolds such as the plane, which go off to infinity. More precisely, I am talking about the distinction between COMPACT [III.9] and noncompact manifolds: a compact manifold can be thought of as one that can be expressed as a closed bounded subset of \mathbb{R}^n for some n . The discussion that follows will be almost entirely about compact manifolds. As some of the examples below will demonstrate, the story for compact manifolds is less convoluted than the analogous story for noncompact ones. For simplicity I shall often use the word “manifold” to mean “compact manifold”; it will be clear from the context if noncompact manifolds are also being discussed.

2.1 Dimension 0

There is only one dimension-0 manifold. It is a single point. The period at the end of this sentence looks, from afar, like a connected, dimension-0 manifold. Note that the distinction between topological and smooth is irrelevant here.

2.2 Dimension 1

There is only one compact, connected, one-dimensional topological manifold, namely the circle. Moreover, the circle has just one smooth structure. Here is one way to represent this structure. Take as a representative circle the unit circle in the xy -plane, that is, the set of all points (x, y) with $x^2 + y^2 = 1$. This can be covered by two overlapping intervals, each of which covers slightly more than half of the circle. The intervals U_1 and U_2 are drawn in figure 3. Each interval constitutes a coordinate chart. The one on the left, U_1 , can be parametrized in a continuous fashion by taking the angle of a given point as measured counterclockwise from the positive x -axis. For example, the point $(1, 0)$ has angle 0, and the point $(-1, 0)$ has angle π . In order to parametrize U_2 by angle, you will have to start with angle π at the negative x -axis. If you move around U_2 , varying this angle continuously, then when you reach the point $(1, 0)$ you will have parametrized it as a point in U_2 using the angle 2π .

As you can see, the arcs U_1 and U_2 intersect in two separated, smaller arcs; these are labeled V_1 and V_2 in

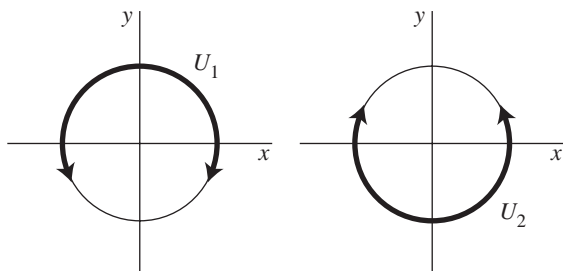


Figure 3 Two charts that cover the circle.

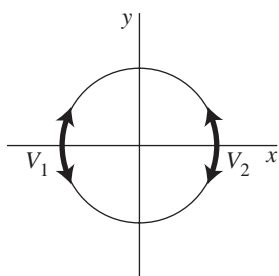
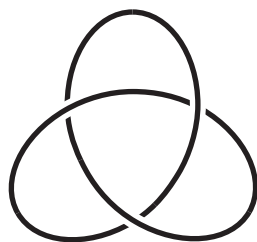
Figure 4 The intersection of the arcs U_1 and U_2 .

Figure 5 A knotted loop in 3-space.

figure 4. The transition function on V_1 is the identity map, since the U_1 angle of any given point in V_1 is the same as its U_2 angle. By contrast, the U_2 angle of a point in V_2 is obtained from the U_1 angle by adding 2π . Thus, the transition function on V_2 is not the identity map but the map that adds 2π to the coordinate function.

This one-dimensional example brings up a number of important issues, all related to a particularly troubling question. To state it, consider first that there are lots of closed loops in the plane that can be taken as model circles. Indeed, the word “lots” considerably understates the situation. Moreover, why should we restrict our attention to circles in the plane? There are closed loops galore in 3-space too: see figure 5, for example. For that matter, any manifold of dimension greater than 1

has smooth loops. Earlier, it was asserted that there is just one smooth, compact, connected, one-dimensional manifold, so all of these loops must be considered the “same.” Why is this?

Here is the answer. We often think of a manifold as it might appear were it sitting in some larger space. For example, we might imagine a circle sitting in the plane, or sitting knotted in three-dimensional Euclidean space. However, the notion of “smooth manifold” introduced above is an *intrinsic* one, in the sense that it does not depend on how the manifold is placed inside a higher-dimensional space. Indeed, it is not even necessary for there to be a higher-dimensional space at all. In the case of the circle, this can be said in the following way. The circle can be placed as a loop in the plane, or as a knot in 3-space, or whatever. Each view of the circle in a higher-dimensional Euclidean space defines a collection of functions that are considered differentiable: one just takes the differentiable functions of the coordinates of the big Euclidean space and restricts them to the circle. As it turns out, any one such collection defines the same smooth structure on the circle as any other. Thus, the smooth structures that are provided by these different views of a circle are all the same, even though there are many interesting ways of placing a circle in a given higher-dimensional space. (In fact, the classification of knots in 3-space is a fascinating, vibrant topic in its own right: see KNOT POLYNOMIALS [III.46].)

How is it proved that there is only one smooth structure for the circle? For that matter, how is it proved that there is but a single compact topological manifold in dimension 1? Since this article is not meant to provide proofs, these questions are left as serious exercises with the following advice. Think hard about the definitions and, for the smooth-manifold question, use some calculus.

2.3 Dimension 2

The story for two-dimensional, connected, compact manifolds is much richer than that for dimension 1. In the first place, there is a basic dichotomy between two kinds of manifold: orientable and nonorientable. Roughly speaking, this is the distinction between manifolds that have two sides and those that have just one. To give a more formal definition, a two-dimensional manifold is called *orientable* if every loop in the manifold that does not cross itself or have any kinks has two distinct sides. This is to say that there is no path from one side of the loop to the other

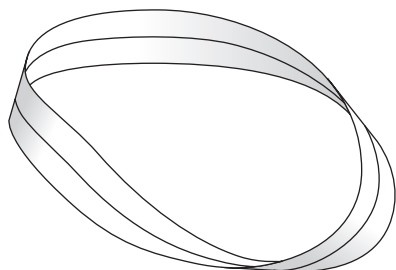


Figure 6 A Möbius strip has just one side.

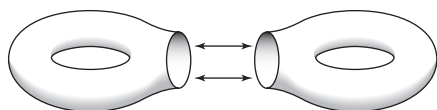


Figure 8 Cutting and gluing.

that avoids the loop yet remains very close to it. The Möbius strip (see figure 6) is not orientable because there are paths from one side of the central loop to the other that do not cross the central loop yet remain very close to it. The orientable, compact, connected, topological, two-dimensional manifolds are in one-to-one correspondence with a collection of fundamental foods: the apple, the doughnut, the two-holed pretzel, the three-holed pretzel, the four-holed pretzel, and so on (see figure 7). Technically, they are classified by an integer, called the *genus*. This is 0 for the sphere, 1 for the torus, 2 for the two-holed torus, etc. The genus counts the number of holes that appear in a given example from figure 7. To say that this classifies them is to say that two such manifolds are the same if and only if they have the same genus. This is a theorem due to POINCARÉ [VI.61].

As it turns out, every topological two-dimensional manifold has exactly one smooth structure, so the list in figure 7 is the same as the list of the *smooth* orientable two-dimensional manifolds. Here one should keep in mind that the notion of a smooth manifold is intrinsic, and therefore independent of how the manifold is represented as a surface in 3-space, or in any other space. For example, the surfaces of an orange, a banana, and a watermelon each represent embedded images of the two-dimensional sphere, the leftmost example in figure 7.

The shapes illustrated in figure 7 suggest an idea that plays a key role when it comes to classifying manifolds of higher dimensions. Notice that the two-holed torus can be viewed as the result of taking two one-holed tori,

cutting disks out of both, gluing the results together across their boundary circles, and then smoothing the corners. This operation is depicted in figure 8. This sort of cutting and gluing operation is an example of what is called a *surgery*. The analogous surgery can also be done with a one-holed torus and a two-holed torus to obtain a three-holed torus. And so on. Thus, all of the oriented two-dimensional manifolds can be built using standard surgeries on copies of just two fundamental building blocks: the one-holed torus and the sphere. Here is a nice exercise to test your understanding of this process. Suppose that you perform a surgery, as in figure 8, on a sphere and another manifold M . Prove that the resulting manifold is the same, with regard to its topological and smooth structure, as M .

As it turns out, all of the nonorientable two-dimensional manifolds can be built using a version of surgery that first cuts a disk out of an orientable two-dimensional manifold and then glues on a Möbius strip. To be more precise, note that the Möbius strip has a circle as its boundary. Cut a disk out of any given orientable, two-dimensional manifold and the result also has a circular boundary. Glue the latter circular boundary to the Möbius strip boundary, smooth the corners, and the result is a smooth manifold that is nonorientable. Every nonorientable, topological (and thus every nonorientable, smooth), two-dimensional manifold is obtained in this way. Moreover, the manifold you get depends only on the number of holes (the genus) of the orientable manifold that is used.

The manifold obtained from the surgery of a Möbius strip with a sphere is called the *projective plane*. The one that uses the Möbius strip and the torus is called the *Klein bottle*. These shapes are illustrated in figure 9. No nonorientable example can be put into three-dimensional Euclidean space in a clean way; any such placement is forced to have portions that pass through other portions, as can be seen in the illustration of the Klein bottle.

How does one prove that the list given above exhausts all two-dimensional manifolds? One method uses versions of the geometric techniques that are discussed below in the three-dimensional context.

2.4 Dimension 3

There is now a complete classification of all smooth, three-dimensional manifolds; this is a very recent achievement. There has been for some time a conjectured list of all three-dimensional manifolds, and a conjectured procedure for telling one from the other. The

PUP: can I double-check that you confirm that the slight leakage of grey on the left-hand side of the two-holed torus is OK? T&T note: ensure that the double-column figure does not appear out of sequence at page make-up stage.

PUP: use of 'corners' is OK here and below.

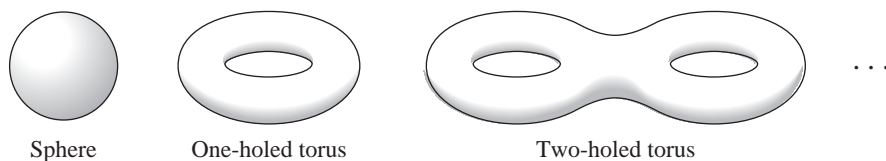


Figure 7 The orientable manifolds of dimension 2.

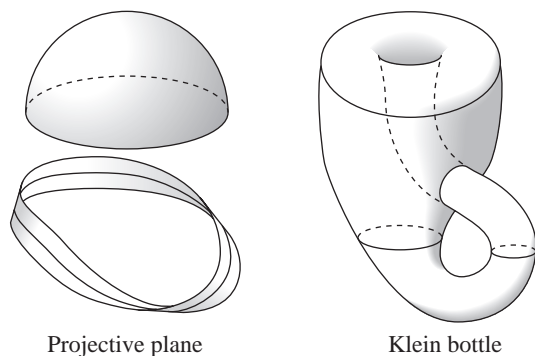


Figure 9 Two nonorientable surfaces. To form the projective plane, one identifies the boundary of the Möbius strip with the boundary of the hemisphere.

proof of these conjectures was recently completed by Grisha Perelman; this is a much-celebrated event in the mathematics community. The proof uses geometry about which more is said in the final part of this article. Here I shall concentrate on the classification scheme.

Before getting to the classification scheme, it is necessary to introduce the notion of a *geometric structure* on a manifold. Roughly speaking, this means a rule for defining the lengths of paths on the manifold. This rule must satisfy the following conditions. The constant path that simply stays at one point has length 0, but any path that moves at all has positive length. Second, if one path starts where another ends, the length of their concatenation (that is, the result of putting the two paths together) is the sum of their lengths.

Note that a rule of this sort for path lengths leads naturally to a notion of distance $d(x, y)$ between any two points x and y on the manifold: one takes the length of the shortest path between them. It turns out to be particularly interesting when $d(x, y)^2$ varies as a smooth function of x and y .

As it happens, there is nothing special about having a geometric structure. Manifolds have them in spades. The following are three very useful geometric structures for the interior of the ball of radius 2 about the origin in n -dimensional Euclidean space. In these for-

mulas, the given path is viewed as if drawn in real time by some hyper-dimensional artist, with $x(t)$ denoting the position of the pencil tip on the path at time t . Here, t ranges over some interval of the real line:

$$\left. \begin{aligned} \text{length} &= \int |\dot{x}(t)| dt; \\ \text{length} &= \int |\dot{x}(t)| \frac{1}{1 + \frac{1}{4}|x(t)|^2} dt; \\ \text{length} &= \int |\dot{x}(t)| \frac{1}{1 - \frac{1}{4}|x(t)|^2} dt. \end{aligned} \right\} \quad (1)$$

In these formulas, \dot{x} denotes the time-derivative of the path $t \rightarrow x(t)$.

The first of these geometric structures leads to the standard Euclidean distance between pairs of points. For this reason it is called the *Euclidean geometry* for the ball. The second defines what is called *spherical geometry* because the distance between any two points is the angle between certain corresponding points in the sphere of radius 1 in $(n + 1)$ -dimensional Euclidean space. The correspondence comes from an $(n + 1)$ -dimensional version of the stereographic projection that is used for maps of the Earth's polar regions. The third distance function defines what is called the *hyperbolic geometry* on the ball. This arises when the ball of radius 2 in n -dimensional Euclidean space is identified in a certain way with a particular hyperbola in $(n + 1)$ -dimensional Euclidean space.

The geometric structures that are depicted in (1) turn out to be symmetrical with respect to rotations and certain other transformations of the unit ball. (You can read more about Euclidean, spherical, and hyperbolic geometry in SOME FUNDAMENTAL MATHEMATICAL DEFINITIONS [I.3 §§6.2, 6.5, 6.6].)

As was remarked above, there are very many geometric structures on any given manifold and so one might hope to find one that has some particularly desirable properties. With this goal in mind, suppose that I have specified some “standard” geometric structure S for the ball in \mathbb{R}^n to serve as a model of an exceptionally desirable structure. This could be one of the ones I have just defined or some other favorite. This leads

to a corresponding notion of the structure S for a compact manifold. Roughly speaking, one says that a geometric structure on a manifold is of the type S if every point in the manifold feels as though it belongs to the unit ball with the structure S , that is, if one can use the structure S on the ball to provide coordinate charts that respect the geometric structure on the manifold. To be more precise, suppose that I am defining a coordinate system in a small neighborhood N of x by means of a function $\phi : N \rightarrow \mathbb{R}^d$. If I can always do this in such a way that the image $\phi(N)$ lies inside the ball, and such that the distance between any two points x and y in N equals the distance between their images $\phi(x)$ and $\phi(y)$, defined in terms of the structure S on the ball, then I will say that the manifold has structure of type S . In particular, a geometric structure is said to be *Euclidean*, *spherical*, or *hyperbolic* when the structure on the ball is Euclidean, spherical, or hyperbolic, respectively.

For example, the sphere in any dimension has a spherical geometric structure (as it should!). As it turns out, every two-dimensional manifold has a geometric structure that is either spherical, Euclidean, or hyperbolic. Moreover, if it has a structure of one of these types, then it cannot have one of a different type. In particular, the sphere has a spherical structure, but not a Euclidean or hyperbolic structure. Meanwhile, the torus in dimension 2 has a Euclidean geometric structure but only a Euclidean one, and all of the other manifolds listed in figure 7 have hyperbolic geometric structures and only hyperbolic ones.

William Thurston had the great insight to realize that three-dimensional manifolds might be classifiable using geometric structures. In particular, he made what was known as the *geometrization conjecture*, which says, roughly speaking, that every three-dimensional manifold is made up of “nice” pieces:

Every smooth three-dimensional manifold can be cut in a canonical fashion along a predetermined set of two-dimensional spheres and one-holed tori so that each of the resulting parts has precisely one of a list of eight possible geometric structures.

The eight possible structures include the spherical, Euclidean, and hyperbolic ones. These plus the other five are, in a sense that can be made precise, those that are maximally symmetric. The other five are associated with various LIE GROUPS [III.50 §1], as are the listed three.

Since its proof by Perelman, the geometrization conjecture has come to be known as the geometrization

theorem. As I shall explain in a moment, this provides a satisfactory resolution of the three-dimensional part of the quest set out at the end of section 1. This is because a manifold with one of the eight geometric structures can be described in a canonical fashion using group theory. As a result, the geometrization theorem turns the classification issue for manifolds into a question that group theory can answer. What follows is an indication of how this comes about.

Each of the eight geometric structures has an associated *model space* which has the given geometric structure. For example, in the case of the spherical structure, the model space is the three-dimensional sphere. For the Euclidean structure, the model space is the three-dimensional Euclidean space. For the hyperbolic structure, it is the hyperbola in the four-dimensional Euclidean space, where the coordinates (x, y, z, t) obey $t^2 = 1 + x^2 + y^2 + z^2$. In all of the eight cases, the model space has a canonical group of self-maps that preserve the distance between any two pairs of points. In the Euclidean case, this group is the group of translations and rotations of the three-dimensional Euclidean space. In the spherical case, it is the group of rotations of the four-dimensional Euclidean space, and in the hyperbolic case, it is the group of Lorentz transformations of four-dimensional Minkowski space. The associated group of self-maps is called the *isometry group* for the given geometric structure.

The connection between manifolds and group theory arises because a certain set of discrete subgroups of the isometry group of any one of the eight model spaces determines a compact manifold with the corresponding geometric structure. (A subgroup is called *discrete* if every point in the subgroup is isolated, meaning that it belongs to a neighborhood that contains no other points from the subgroup.) This compact manifold is obtained as follows. Two points x and y in the model space are declared to be *equivalent* if there is an isometry T , belonging to the subgroup, such that $Tx = y$. In other words, x is equivalent to all its images under isometries from the subgroup. It is easy to check that this notion of equivalence is a genuine EQUIVALENCE RELATION [I.2 §2.3]. The equivalence classes are then in one-to-one correspondence with the points of the associated compact manifold.

Here is a one-dimensional example of how this works. Think of the real line as a model space whose isometry group is the group of translations. The set of translations by integer multiples of 2π forms a discrete subgroup of this group. Given a point t in the real line, the

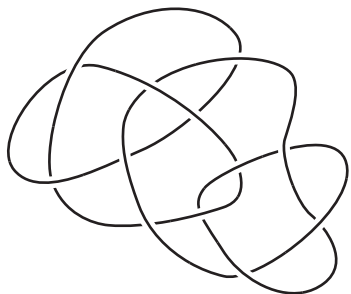


Figure 10 A link formed out of two knots.

possible images under translations from the subgroup are all the numbers of the form $t + 2n\pi$, where n is an integer, so one regards two real numbers as equivalent if they differ by a multiple of 2π , and the equivalence class of t is $\{t + 2n\pi : n \in \mathbb{Z}\}$. One can associate with this equivalence class the point $(x, y) = (\cos t, \sin t)$ in the circle, since adding a multiple of 2π to t does not affect either its sine or its cosine. (Intuitively speaking, if you regard each t as equivalent to $t + 2\pi$, then you are wrapping the real line around and around a circle.)

This association between certain subgroups of the isometry group and compact manifolds with the given geometric structure goes in the other direction as well. That is, the subgroup can be recovered from the manifold in a relatively straightforward fashion using the fact that each point in the manifold lies in a coordinate chart where its distance function is the same as that of the associated model space.

Even before Perelman's work there was a tremendous amount of evidence for the validity of the geometrization conjecture, much of it supplied by Thurston. In order to discuss this evidence, a small digression is required to give some of the background. First, I need to bring in the notion of a *link* in the three-dimensional sphere. A link is the name given to a finite disjoint union of knots. Figure 10 depicts an example of one that is made out of two knots.

I also need the notion of *surgery on a link*. To this end, thicken the link so as to view it as a union of knotted, solid tubes. (Think of the knot as the copper in an insulated wire and view the solid tube as the copper plus the surrounding insulation.) Notice that the boundary of any given component tube is really a copy of our one-holed torus from figure 7. Therefore, removing any one of the tubes leaves a tubular-shaped missing region from the three-dimensional sphere whose boundary is a torus.

Now, to define a surgery, imagine removing a knotted tube and then gluing it back in a different way. That is, imagine gluing the boundary of the tube to the boundary of the resulting missing region using an identification that is *not* the same as the original. For example, take the “unknot,” a standard round circle in a given plane, here viewed as living inside a coordinate chart of the three-dimensional sphere. Take out the solid tube around it, and then replace the tube by gluing the boundary in the “wrong” way, as follows. Consider the leftmost torus in figure 11 as the boundary of the complement of the tube in \mathbb{R}^3 . Consider the middle torus as the inside of the tube. The “wrong” gluing identifies the circles marked “R” and “L” on the leftmost torus with their counterparts on the middle torus. The resulting space is a three-dimensional manifold which turns out to be the product of the circle with the two-dimensional sphere. That is to say, it is the set of ordered pairs (x, y) , where x is a point in the circle and y is a point in the two-dimensional sphere. There are many other possible ways to glue the boundary torus, and almost all of the corresponding surgeries give rise to distinct three-dimensional manifolds. One of these is illustrated in the rightmost part of figure 11.

In general, given any link one can construct a countably infinite set of distinct, smooth three-dimensional manifolds by using surgeries on it. Furthermore, Raymond Lickorish proved that *every* three-dimensional manifold can be obtained by using surgery on *some* link in the three-dimensional sphere. Unfortunately, this characterization of three-dimensional manifolds via surgeries on links does not provide a satisfactory resolution to the central quest of classifying smooth structures because the process is far from unique: for any given manifold there is a bewildering assortment of links and surgeries that can be used to produce it. Moreover, as of this writing, there is no known way to classify knots and links in the three-dimensional sphere.

In any event, here is a taste of Thurston's evidence for his geometrization conjecture. Given any link, all but finitely many of the three-dimensional manifolds you can produce from it by surgery satisfy the conclusions of the geometrization conjecture. Thurston also proved that, given any knot apart from the unknot, all but finitely many surgeries on it produce a manifold with a hyperbolic geometric structure.

By the way, Perelman's proof of the geometrization theorem gives as a special case a proof of the *Poincaré conjecture*, proposed by Poincaré in 1904. To state this

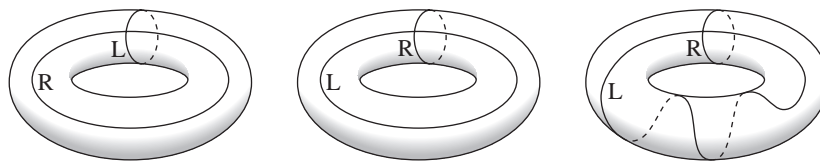


Figure 11 Different ways of gluing a tube into a tube-shaped hole.

we need the notion of a *simply connected* manifold. This is a manifold with the property that any closed loop in it can be shrunk down to a point. To be more precise, designate a point in the manifold as the “base point.” Then any path in the manifold that starts and ends at the chosen base point can be continuously deformed in such a way that at each stage of the deformation the path still starts and ends at the base point, and so that the end result is the trivial path that starts at the base point and just stays there. For example, the two-dimensional sphere is simply connected, but the torus is not, since a loop that goes “once around” the torus (for example, any of the loops R or L in the various tori of figure 11) cannot be shrunk to a point. In fact, a sphere is the only two-dimensional manifold that is simply connected, and spheres are simply connected in all dimensions greater than 1.

The Poincaré conjecture. Every compact, simply connected, three-dimensional manifold is the three-dimensional sphere.

2.5 Dimension 4

This is the weird dimension. Nobody has managed to formulate a useful and viable conjecture for the classification of smooth, compact, four-dimensional manifolds. On the other hand, the classification story for many categories of topological four-dimensional manifolds is well-understood. For the most part, this work is by Michael Freedman.

Some of the topological manifolds in dimension 4 do not admit smooth structures. The so-called “ $\frac{11}{8}$ conjecture” proposes necessary and sufficient conditions for a four-dimensional, topological manifold to have at least one smooth structure. The fraction $\frac{11}{8}$ here refers to the absolute value of the ratio of the rank to the signature of a certain symmetric, bilinear form that appears in the four-dimensional story. The case $\frac{0}{0}$ excepted, the conjecture asserts that a smooth structure exists if and only if this ratio is at least $\frac{11}{8}$. The bilinear form in question is obtained by counting with

signed weights the intersection points between various two-dimensional surfaces inside the given four-dimensional manifold. In this regard, note that a typical pair of two-dimensional surfaces in four dimensions will intersect at finitely many points. This is a higher-dimensional analogue of a fact that is rather easier to visualize: that a typical pair of loops in the two-dimensional plane will intersect at finitely many points. Not surprisingly, the bilinear form here is called the *intersection form*; it plays a prominent role in Freedman’s classification theorems.

Meanwhile, the problem of listing all smooth structures is wide open in four dimensions: there are no cases of a topological manifold with at least one smooth structure where the list of distinct structures is known to be complete. Some topological four-dimensional manifolds are known to have (countably) infinitely many distinct smooth structures. For others there is only one known structure. For example, the four-dimensional sphere has one obvious smooth structure and this is the only one known. However, the underlying topological manifold may, for all anyone knows, have many distinct smooth structures. By the way, the story for noncompact manifolds in dimension 4 is truly bizarre. For example, it is known that there are uncountably many smooth manifolds that are homeomorphic to the standard, four-dimensional Euclidean space. But even here, our understanding is less than optimal since there is no known explicit construction of a single one of these “exotic” smooth structures.

Simon Donaldson provided a set of geometric invariants that have the power to distinguish smooth structures on a given topological 4-manifold. Donaldson’s invariants were recently superseded by a suite of more computable invariants; these were proposed by Edward Witten and are called the *Seiberg–Witten invariants*. More recently still, Peter Ozsvath and Zoltan Szabo designed a possibly equivalent set of invariants that are even easier to use. Do the Seiberg–Witten invariants (broadly defined) distinguish all smooth structures? No

one knows. A bit more is said about these invariants in the final part of this article.

Note that Freedman's results include the topological version of the four-dimensional Poincaré conjecture that follows.

The four-dimensional sphere is the only compact, topological 4-manifold with the following property: every based map from either a one-dimensional circle or a two-dimensional sphere can be continuously deformed so that the result maps onto the base point.

The smooth version of this conjecture has not been resolved.

Is there a four-dimensional version of the geometrization conjecture/theorem?

2.6 Dimensions 5 and Greater

Surprisingly enough, the issues raised at the end of the first section have more or less been resolved in all dimensions that are greater than 4. This was done some time ago by Stephen Smale with input from John Stallings. In these higher dimensions it is also possible to say what conditions need to hold in order for a topological manifold to admit a smooth structure. For example, John Milnor and others determined that the respective number of smooth structures on the spheres of dimensions 5–18 are as follows: 1, 1, 28, 2, 8, 6, 992, 1, 3, 2, 16 256, 2, 16, 16.

At first sight, it is surprising that the dimensions greater than 4 are easier to deal with than dimensions 3 and 4. However, there is a good reason for this. It turns out that there is more room to maneuver in these higher-dimensional spaces and this extra room makes all the difference. To get a sense for this, let n be a positive integer, and let S^n denote the n -dimensional sphere. To make this more concrete, view S^n as the set of points (x_1, \dots, x_{n+1}) in the Euclidean space \mathbb{R}^n such that $x_1^2 + \dots + x_{n+1}^2 = 1$. Now consider the product manifold, $S^n \times S^n$. This is the set of pairs of points (x, y) , where x is in one copy of S^n and y is in another. This product manifold has dimension $2n$. A standard picture of $S^n \times S^n$ has two distinguished copies of S^n inside it, one consisting of all points of the form (x, y) with $y = (1, 0, \dots)$ and the other consisting of all points (x, y) with $x = (1, 0, \dots)$. Let us call the first copy S_R and the second one S_L . Of particular interest here is the fact that S_R and S_L intersect in precisely one point, the point $((1, 0, \dots), (1, 0, \dots))$.

By the way, in the $n = 1$ case, the space $S^1 \times S^1$ is the doughnut in figure 7. The one-dimensional spheres

S_R and S_L inside it are the circles that are drawn in the leftmost diagram in figure 11.

If you are with me so far, suppose now that an advanced alien en route from Arcturus to the galactic center kidnaps you and drops you into some unknown, $2n$ -dimensional manifold. You suspect that it is $S^n \times S^n$, but are not sure. One reason that you suspect this to be the case is that you have found a pair of n -dimensional spheres in it, one you call M_R and the other you call M_L . Unfortunately, they intersect in $2N + 1$ points, where $N > 0$. You would be less nervous about things if you could find a pair of different spheres that intersect precisely once. So you wonder whether perhaps you can push M_L around a bit so as to remove the $2N$ unwanted intersection points.

The surprise here is that the issue of removing intersection points in any dimension concerns only certain zero-, one-, and two-dimensional manifolds that live inside your $2n$ -dimensional one. This is an old observation due to Hassler Whitney. In particular, Whitney discovered that in the $2n$ -dimensional manifold you must be able to find a disk of dimension two whose boundary loop lies half in M_L and half in M_R . This boundary loop must hit two of the intersection points (one when it passes from M_L to M_R and one when it passes back again). The disk must also stick out orthogonally to M_L and M_R where it touches them. If its interior is disjoint from both M_L and M_R , and if there are no points where the disk comes back to intersect itself, then you can push the part of M_L that is very near the disk along the disk while stretching the remaining part to keep things from tearing. If you extend the disk a bit past M_R , then you will have removed two of the intersection points when you have pushed past the end of the disk. Figure 12 is a schematic of this. This pushing operation (the *Whitney trick*) can be performed in any manifold of any dimension if you can find the required disk. The problem is to find the disk. Figure 13 is a drawing of a cross-sectional slice showing a "good" disk on the left and some badly chosen disks in the middle and on the right. If you have a badly chosen disk that nevertheless satisfies the required boundary conditions, then you might hope to find a tiny wiggle of its interior that makes it better. You would like the new disk to have no self-intersection points and you would like its interior to be disjoint from both M_L and M_R . No wiggle along a direction that is parallel to the disk itself will help, for any such wiggle only changes the position of the intersection point in the disk. Likewise, a wiggle in a direction parallel to the offending M_L or M_R is useless since

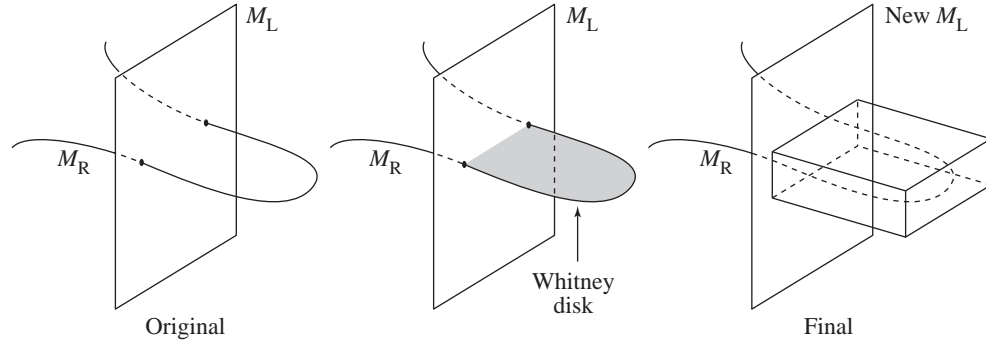


Figure 12 The Whitney trick.

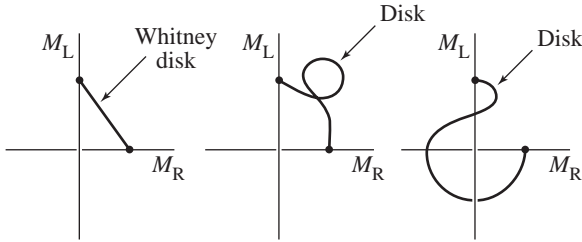


Figure 13 Some possible Whitney disks.

it only changes the position of the intersection point in the latter space. Thus, $2 + n$ of the $2n$ dimensions are useless when it comes to wiggling a disk. However, there are $2n - (n + 2) = n - 2$ remaining dimensions to work with, which is a positive number when $2n > 4$. In fact, when this is true a generic wiggle in any of these extra dimensions does the trick.

Now, when $2n = 4$ (so $n = 2$) there are no extra dimensions, and, consequently, no small wiggle can make a new disk without intersection points. So if a given candidate disk intersects M_R , then the Whitney trick just trades the old pair of intersection points for a new collection. If the disk intersects either itself or M_L , then the new version of M_L has self-intersection points: that is, points where one part has come around to intersect another.

This failure of the Whitney trick is the bane of four-dimensional topology. Thus, a major lemma for Michael Freedman's classification theorem about topological four-dimensional manifolds describes ubiquitous circumstances where a topologically (but not smoothly!) embedded disk can be found for use in the Whitney trick.

3 How Geometry Enters the Fray

Much of our current understanding about smooth manifolds in dimensions 4 or less has come via what might be called geometric techniques. The search for a canonical geometric structure on a given three-dimensional manifold is an example. Perelman's proof of the geometrization theorem proceeds in this manner. The idea is to choose any convenient geometric structure on a given three-dimensional manifold and then continuously deform it by some well-defined rule. If one views the deformation as a time-dependent process, then the goal is to design the deformation rule to make the geometric structure ever more symmetric as time goes on.

A rule introduced and much studied by Richard Hamilton and then used by Perelman specifies the time-derivative of the geometric structure at any given time in terms of certain of its properties at that time. It is a nonlinear version of the classical HEAT EQUATION [I.3 §5.4]. For those unfamiliar with the latter, the simplest version modifies functions on the real line and will now be described. Let τ denote the time parameter, and let $f(x)$ denote a given function on the line, representing the initial distribution of heat. The resulting time-dependent family of functions associates with any given positive value for τ a function, $F_\tau(x)$, which represents the distribution of heat at time τ . The partial derivative of $F_\tau(x)$ with respect to τ is equal to its second partial derivative with respect to x , and the initial condition is that $F_0(x) = f(x)$. If the initial function f is zero outside some interval, then one can write down a formula for F_τ :

$$F_\tau(x) = \frac{1}{(2\pi\tau)^{1/2}} \int_{-\infty}^{\infty} e^{-(x-y)^2/2\tau} f(y) dy. \quad (2)$$

One can see from (2) that $F_\tau(x)$ tends uniformly to zero in x as τ tends to infinity. In particular, this limit is completely ignorant of the starting function f ; and, being identically zero, it is also the most symmetric function possible. The representation for F_τ in (2) indicates how this comes about. The value of F_τ at any given point is a weighted average of the values of the original function. Moreover, as τ increases, this average looks more like the standard average over ever-larger regions of the line. Physically this is very plausible as well: the heat spreads itself out more and more thinly as time goes on.

The time-dependent family of geometric structures that Hamilton introduced and Perelman used is defined by an equation that relates the time-derivative of the geometric structure at any given time to its *Ricci curvature*, a certain natural substitute in the context of geometric structures for the second derivatives that enter the heat equation for the functions F_τ above. The idea much studied by Hamilton and then by Perelman is to let the evolving geometric structure decompose the manifold into the canonical pieces that are predicted to exist by the geometrization conjecture. Perelman proved that the pieces required by the geometrization conjecture emerge as regions whose points stay relatively close together (as measured by a certain rescaling of the distance function) while the points in distinct regions move farther and farther apart.

The equation used by Perelman and Hamilton for the time-evolution of a geometric structure is rather complicated. Its standard incarnation involves the notion of a *RIEMANNIAN METRIC* [I.3 §6.10]. This appears in any given coordinate chart on an n -dimensional manifold as a symmetric, positive-definite $n \times n$ matrix whose entries are functions of the coordinates. The various components of this matrix are traditionally written as $\{g_{ij}\}_{1 \leq i, j \leq n}$. The matrix determines the geometric structure and can in turn be derived from it.

Hamilton and Perelman study a time-dependent family of Riemannian metrics, $\tau \rightarrow g_\tau$, where the rule for the time dependence is obtained using an equation for the τ -derivative of g_τ that has the schematic form $\partial_\tau(g_\tau)_{ij} = -2R_{ij}[g_\tau]$, where $\{R_{ij}\}_{1 \leq i, j \leq n}$ are the components of the aforementioned Ricci curvature, a certain symmetric matrix that is determined at any given τ by the metric g_τ . Every Riemannian metric has a Ricci curvature; its components are standard (nonlinear) functions of the components of the matrix and their first- and second-order partial derivatives in the coordinate directions. The Ricci curvatures for the

metrics that define the respective Euclidean, spherical, and hyperbolic geometries have the particularly simple form $R_{ij} = c g_{ij}$, where c is 0, 1, or -1 , respectively. For more about these ideas, see RICCI FLOW [III.80].

As was mentioned at the beginning of this part of the article, geometry has also played a central role in the developments in the classification program for smooth, four-dimensional manifolds. In this case, geometrically defined data are used to distinguish smooth structures on topologically equivalent manifolds. What follows is a very brief sketch of how this is done.

To begin with, the idea is to introduce a geometric structure on the manifold and then to use the latter to define a canonical system of partial differential equations. In any given coordinate chart, these equations are for a particular set of functions. The equations state that certain linear combinations of the collection of first derivatives of the functions from the set are equal to terms that are linear and quadratic in the values of the functions themselves. In the case of the Donaldson invariants, and also of the newer Seiberg-Witten invariants, the relevant equations are nonlinear generalizations of the MAXWELL EQUATIONS [IV.13 §1.1] for electricity and magnetism.

In any event, one then counts the solutions with algebraic weights. The purpose of the algebraic weighting of the count is to obtain an INVARIANT [I.4 §2.2], that is, a count that does not change if the given geometric structure is changed. The point here is that the naive count will typically depend on the structure, but a suitably weighted count will not. Imagine, for example, that one has a continuously varying family of geometric structures, and that new solutions appear and old ones disappear only in pairs, where one solution has been assigned weight $+1$ and the other -1 .

The following toy model illustrates this appearance and disappearance phenomenon. The equation in question is for a single function on the circle. That is, it will concern a function, f , of one variable, x , that is periodic with period 2π . For example, take the equation $\partial f / \partial x + \tau f - f^3 = 0$, where τ is a constant that is specified in advance. Varying τ can now be viewed as a model for the variation of the geometric structure. When $\tau > 0$ there are exactly three solutions: $f \equiv 0$, $f \equiv \tau$, and $f \equiv -\tau$. However, when $\tau \leq 0$, the only solution is $f \equiv 0$. Thus, the number of solutions changes as τ crosses zero. Even so, a suitable weighted count is independent of τ .

Let us return now to the four-dimensional story. If the weighted sum is independent of the chosen *geo-*

PUP: OK that this author definitely wanted to keep his use of 'data' as the plural use, whereas in most places we are using it as a singular noun?

T&T note: need to check all index entries to 'geometric structure' as some of them should be 'geometric structure on a manifold', some should be 'geometric structure on a group', and so on.

metric structure, then it depends only on the underlying *smooth* structure. Therefore, if two geometric structures on a given topological manifold provide distinct sums, then the underlying smooth structures must be distinct.

As I remarked earlier, Ozsvath and Szabo have defined invariants for four-dimensional manifolds that are easier to use than the Seiberg–Witten invariants, but probably equivalent to them. These are also defined as the number of solutions to a particular system of differential equations, counted in a creative way. In this case, the equations are analogues of the CAUCHY-RIEMANN EQUATIONS [I.3 §5.6], and the arena is a space that can be defined after cutting the 4-manifold into simpler pieces. There are myriad ways to slice a 4-manifold in the prescribed manner, but a suitably creative, algebraic count of solutions provides the same number for each.

With hindsight, one can see that the use of differential equations to distinguish smooth structures on a given topological manifold makes good sense, since a smooth structure is needed to take a derivative in the first place. Even so, this author is constantly amazed by the fact that the Donaldson/Seiberg–Witten/Ozsvath–Szabo strategy of algebraically counting differential equation solutions yields counts that are both tractable and useful. (Getting the same count in all cases is no help at all.)

Further Reading

Those who wish to learn more about manifolds in general can consult J. Milnor’s book *Topology from the Differentiable Viewpoint* (Princeton University Press, Princeton, NJ, 1997) or the book *Differential Topology* (Prentice Hall, Englewood Cliffs, NJ, 1974), by V. Guillemin and A. Pollack. A good introduction to the classification problem in dimensions 2 and 3 is the book *Three-Dimensional Geometry and Topology* (Princeton University Press, Princeton, NJ, 1997), by W. Thurston. This book also has a nice discussion of geometric structures. A full account of Perelman’s proof of the Poincaré conjecture can be found in *Ricci Flow and the Poincaré Conjecture*, by J. Morgan and G. Tian (American Mathematical Society, Providence, RI, 2007). The story for topological 4-manifolds is told in the book by M. Freedman and F. Quinn titled *Topology of 4-Manifolds* (Princeton University Press, Princeton, NJ, 1990). There are no books available that serve as general introductions to the smooth 4-manifold story.

A book that does introduce the Seiberg–Witten invariants is *The Seiberg–Witten Equations and Applications to the Topology of Smooth Four-Manifolds* (Princeton University Press, Princeton, NJ, 1995), by J. Morgan. Meanwhile, the Donaldson invariants are discussed in detail in the book by Donaldson and P. Kronheimer titled *Geometry of Four-Manifolds* (Oxford University Press, Oxford, 1990). Finally, parts of the story for dimensions greater than 4 are told in *Lectures on the h-Cobordism Theorem* (Princeton University Press, Princeton, NJ, 1965), by J. Milnor, and *Foundational Essays on Topological Manifolds, Smoothings and Triangulations* (Princeton University Press, Princeton, NJ, 1977), by R. Kirby and L. Siebenman.

IV.8 Moduli Spaces

David D. Ben-Zvi

Many of the most important problems in mathematics concern CLASSIFICATION [I.4 §2]. One has a class of mathematical objects and a notion of when two objects should count as equivalent. It may well be that two equivalent objects look superficially very different, so one wishes to describe them in such a way that equivalent objects have the same description and inequivalent objects have different descriptions.

Moduli spaces can be thought of as *geometric* solutions to *geometric* classification problems. In this article we shall illustrate some of the key features of moduli spaces, with an emphasis on the moduli spaces of RIEMANN SURFACES [III.81]. In broad terms, a *moduli problem* consists of three ingredients.

Objects: which geometric objects would we like to describe, or *parametrize*?

Equivalences: when do we identify two of our objects as being isomorphic, or “the same”?

Families: how do we allow our objects to vary, or modulate?

In this article we will discuss what these ingredients signify, as well as what it means to *solve* a moduli problem, and we will give some indications as to why this might be a good thing to do.

Moduli spaces arise throughout ALGEBRAIC GEOMETRY [IV.4], differential geometry, and ALGEBRAIC TOPOLOGY [IV.6]. (Moduli spaces in topology are often referred to as *classifying spaces*.) The basic idea is to give a geometric structure to the *totality* of the objects we are trying to classify. If we can understand this geometric structure, then we obtain powerful insights into

the geometry of the objects themselves. Furthermore, moduli spaces are rich geometric objects in their own right. They are “meaningful” spaces, in that any statement about their geometry has a “modular” interpretation, in terms of the original classification problem. As a result, when one investigates them one can often reach much further than one can with other spaces. Moduli spaces such as the moduli of ELLIPTIC CURVES [III.21] (which we discuss below) play a central role in a variety of areas that have no immediate link to the geometry being classified, in particular in ALGEBRAIC NUMBER THEORY [IV.1] and algebraic topology. Moreover, the study of moduli spaces has benefited tremendously in recent years from interactions with physics (in particular with STRING THEORY [IV.17 §2]). These interactions have led to a variety of new questions and new techniques.

1 Warmup: The Moduli Space of Lines in the Plane

Let us begin with a problem that looks rather simple, but that nevertheless illustrates many of the important ideas of moduli spaces.

Problem. Describe the collection of all lines in the real plane \mathbb{R}^2 that pass through the origin.

To save writing, we are using the word “line” to mean “line that passes through the origin.” This classification problem is easily solved by assigning to each line L an essential parameter, or *modulus*, a quantity that we can calculate for each line and that will help us tell different lines apart. All we have to do is take standard Cartesian coordinates x, y on the plane and measure the angle $\theta(L)$ between the line L and the x -axis, taken in counterclockwise fashion. We find that the possible values of θ are those for which $0 \leq \theta < \pi$, and that for every such θ there is exactly one line L that makes an angle of θ with the x -axis. So as a *set*, we have a complete solution to our classification problem: the set of lines L , known as *the real projective line* \mathbb{RP}^1 , is in one-to-one correspondence with the half-open interval $[0, \pi)$.

However, we are seeking a *geometric* solution to the classification problem. What does this entail? We have a natural notion of when two lines are near each other, which our solution should capture—in other words, the collection of lines has a natural TOPOLOGY [III.92]. So far, our solution does not reflect the fact that lines L for which the angle $\theta(L)$ is close to π are almost horizontal: they are therefore close to the x -axis (for

which $\theta = 0$) and to the lines L with $\theta(L)$ close to zero. We need to find some way of “wrapping around” the interval $[0, \pi)$ so that π becomes close to 0.

One way to do this is to take not the half-open interval $[0, \pi)$ but the closed interval $[0, \pi]$, and then to “identify” the points 0 and π . (This idea can easily be made formal by defining an appropriate EQUIVALENCE RELATION [I.2 §2.3].) If π and 0 are regarded as the same, then numbers close to π are close to numbers close to 0. This is a way of saying that if you attach the two ends of a line segment together, then, topologically speaking, you obtain a circle.

A more natural way of achieving the same end is suggested by the following geometric construction of \mathbb{RP}^1 . Consider the unit circle $S^1 \subset \mathbb{R}^2$. To each point $s \in S^1$, there is an obvious way of assigning a line $L(s)$: take the line that passes through s and the origin. Thus, we have a *family of lines parametrized by* S^1 , that is, a map (or function) $s \mapsto L(s)$ that takes points in S^1 to lines in our set \mathbb{RP}^1 . What is important about this is that we already know what it means for two points in S^1 to be close to each other, and the map $s \mapsto L(s)$ is continuous. However, this map is a two-to-one function rather than a bijection, since s and $-s$ always give the same line. To remedy this, we can identify each s in the circle S^1 with its antipodal point $-s$. We then have a one-to-one correspondence between \mathbb{RP}^1 and the resulting QUOTIENT SPACE [I.3 §3.3] (which again is topologically a circle), and this correspondence is continuous in both directions.

The key feature of the space \mathbb{RP}^1 , considered as the *moduli space* of lines in the plane, is that it captures the ways in which lines can *modulate*, or vary continuously in families. But when do families of lines arise? A good example is provided by the following construction. Whenever we have a continuous curve $C \subset \mathbb{R}^2 \setminus 0$ in the plane, we can assign to each point c in C the line $L(c)$ that passes through 0 and c . This gives us a family of lines parametrized by C . Moreover, the function that takes c to $L(c)$ is a continuous function from C to \mathbb{RP}^1 , so the parametrization is a continuous one.

Suppose, for example, that C is a copy of \mathbb{R} realized as the set of points $(x, 1)$ at height 1. Then the map from C to \mathbb{RP}^1 gives an isomorphism between \mathbb{R} and the set $\{L : \theta(L) \neq 0\}$, which is the subset of \mathbb{RP}^1 consisting of all lines apart from the x -axis. Put more abstractly, we have an intuitive notion of what it means for a collection of lines through the origin to depend continuously on some parameters, and this notion is captured precisely by the geometry of \mathbb{RP}^1 : for instance, if you

PUP: this is a different object from ‘ \mathbb{RP}^1 ’ discussed below. All text is OK.

tell me you have a continuous 37-parameter family of lines in \mathbb{R}^2 , this is the same as saying that you have a continuous map from \mathbb{R}^{37} to \mathbb{RP}^1 , which sends a point $v \in \mathbb{R}^{37}$ to a line $L(v) \in \mathbb{RP}^1$. (More concretely, we could say that the real function $v \mapsto \theta(L(v))$ on \mathbb{R}^{37} is continuous away from the locus where θ is close to π . Near this locus we could use instead the function ϕ that measures the angle from the y -axis.)

1.1 Other Families

The idea of families of lines leads to various other geometric structures on the space \mathbb{RP}^1 , and not just its topological structure. For example, we have the notion of a *differentiable* family of lines in the plane, which is a family of lines for which the angles vary differentiably. (The same ideas apply if we replace “differentiable” by “measurable,” “ C^∞ ,” “real analytic,” etc.) To parametrize such a family appropriately, we would like \mathbb{RP}^1 to be a DIFFERENTIABLE MANIFOLD [L3 §6.9], so that we can calculate derivatives of functions on it. Such a structure on \mathbb{RP}^1 can be specified by using the angle functions θ and ϕ defined in the previous section. The function θ gives us a coordinate for lines that are not too close to the x -axis, and ϕ gives us a coordinate for lines that are not too close to the y -axis. We can calculate derivatives of functions on \mathbb{RP}^1 by writing them in terms of these coordinates. One can justify this differentiable structure on \mathbb{RP}^1 by checking that for any differentiable curve $C \subset \mathbb{R}^2 \setminus 0$ the map $c \mapsto L(c)$ comes out as differentiable. This means that if $L(c)$ is not close to the x -axis, then the function $x \mapsto \theta(L(x))$ is differentiable at $x = c$, and similarly for ϕ and the y -axis. The functions $x \mapsto \theta(L(x))$ and $x \mapsto \phi(L(x))$ are called *pullbacks*, because they are the result of converting, or “pulling back,” θ and ϕ from functions defined on \mathbb{RP}^1 to functions defined on C .

We now can state the fundamental property of \mathbb{RP}^1 as a differentiable space.

A differentiable family of lines in \mathbb{R}^2 parametrized by a differentiable manifold X is the same thing as a function from X to \mathbb{RP}^1 , taking a point x to a line $L(x)$, such that the pullbacks $x \mapsto \theta(L(x))$ and $x \mapsto \phi(L(x))$ of the functions θ, ϕ are differentiable functions.

We say that \mathbb{RP}^1 (with its differentiable structure) is the *moduli space* of (differentiably varying families of) lines in \mathbb{R}^2 . This means that \mathbb{RP}^1 carries the *universal differentiable family of lines*. From the very definition, we have assigned to each point of \mathbb{RP}^1 a line in \mathbb{R}^2 , and

these lines vary differentiably as we vary the point. The above assertion says that *any* differentiable family of lines, parametrized by a space X , is described by giving a map $f : X \rightarrow \mathbb{RP}^1$ and assigning to $x \in X$ the line $L(f(x))$.

1.2 Reformulation: Line Bundles

It is interesting to reformulate the notion of a (continuous or differentiable) family of lines as follows. Let X be a space and let $x \mapsto L(x)$ be an assignment of lines to points in X . For each point $x \in X$, we place a copy of \mathbb{R}^2 at x ; in other words, we consider the Cartesian product $X \times \mathbb{R}^2$. We may now visualize the line $L(x)$ as living in the copy of \mathbb{R}^2 that lies over x . This gives us a continuously varying collection of lines $L(x)$ parametrized by $x \in X$, otherwise known as a *line bundle* over X . Moreover, this line bundle is embedded in the “trivial” VECTOR BUNDLE [IV.6 §5] $X \times \mathbb{R}^2$, which is the constant assignment that takes each x to the plane \mathbb{R}^2 . In the case when X is \mathbb{RP}^1 itself, we have a “tautological” line bundle: to each point $s \in \mathbb{RP}^1$, which we can think of as a line L_s in \mathbb{R}^2 , it assigns that very same line L_s .

Proposition. *For any topological space X there is a natural bijection between the following two sets:*

- (i) *the set of continuous functions $f : X \rightarrow \mathbb{RP}^1$; and*
- (ii) *the set of line bundles on X that are contained in the trivial vector bundle $X \times \mathbb{R}^2$.*

This bijection sends a function f to the corresponding pullback of the tautological line bundle on \mathbb{RP}^1 . That is, the function f is mapped to the line bundle $x \mapsto L_{f(x)}$. (This is a pullback because it converts L from a function defined on \mathbb{RP}^1 to a function defined on X .)

Thus, the space \mathbb{RP}^1 carries the *universal* line bundle that sits in the trivial \mathbb{R}^2 bundle—any time we have a line bundle sitting in the trivial \mathbb{R}^2 bundle, we can obtain it by pulling back the universal (tautological) example on \mathbb{RP}^1 .

1.3 Invariants of Families

Associated with any continuous function f from the circle S^1 to itself is an integer known as its *degree*. Roughly speaking, the degree of f is the number of times $f(x)$ goes around the circle when x goes around once. (If it goes backwards n times, then we say that the degree is $-n$.) Another way to think of the degree is as the number of times a typical point in S^1 is passed by

$f(x)$ as x goes around the circle, where we count this as $+1$ if it is passed in the counterclockwise direction and -1 if it is passed in the clockwise direction.

Earlier, we showed that the circle S^1 , which we obtained by identifying the endpoints of the closed interval $[0, \pi]$, could be used to parametrize the moduli space \mathbb{RP}^1 of lines. Combining this with the notion of degree, we can draw some interesting conclusions. In particular, we can define the notion of *winding numbers*. Suppose that we are given a continuous function γ from the circle S^1 into the plane \mathbb{R}^2 and suppose that it avoids 0. The image of this map will be a closed loop C (which may cross itself). This defines for us a map from S^1 to itself: first do γ to obtain a point c in C , then work out $L(c)$, which belongs to \mathbb{RP}^1 , and finally use the parametrization of \mathbb{RP}^1 to associate with $L(c)$ a point in S^1 again. The degree of the resulting composite map will be *twice* the number of times that γ , and hence C , winds around 0, so half this number is defined to be the winding number of γ .

More generally, given a family of lines in \mathbb{R}^2 parametrized by some space X , we would like to measure the “manner in which X winds around the circle.” To be precise, given a function ϕ from X to \mathbb{RP}^1 , which defines the parametrized family of lines, we would like to be able to say, for any map $f : S^1 \rightarrow X$, what the winding number is of the composition ϕf , which takes a point x in S^1 to its image $f(x)$ in X and from there to the corresponding line $\phi(f(x))$ in the family. Thus, the map ϕ gives us a way of assigning to each function $f : S^1 \rightarrow X$ an integer, the winding number of ϕf . The way this assignment works does not change if ϕ is continuously deformed: that is, it is a topological invariant of ϕ . What it does depend on is the class that ϕ belongs to in the first COHOMOLOGY GROUP [IV.6 §4] of X , $H^1(X, \mathbb{Z})$. Equivalently, to any line bundle on a space X which is contained in the trivial \mathbb{R}^2 -bundle, we have associated a cohomology class, known as the *Euler class* of the bundle. This is the first example of a CHARACTERISTIC CLASS [IV.6 §5] for vector bundles. It demonstrates that if we understand the topology of moduli spaces of classes of geometric objects, then we can define topological invariants for families of those objects.

2 The Moduli of Curves and Teichmüller Spaces

We now turn our attention to perhaps the most famous examples of moduli spaces, the moduli spaces of

curves, and their first cousins, the *Teichmüller spaces*. These moduli spaces are the geometric solution to the problem of classification of compact Riemann surfaces, and can be thought of as the “higher theory” of Riemann surfaces. The moduli spaces are “meaningful spaces,” in that each of their points stands for a Riemann surface. As a result, any statement about their geometry tells us something about the geometry of Riemann surfaces.

We turn first to the objects. Recall that a *Riemann surface* is a topological surface X (connected and oriented) to which a *complex structure* has been given. Complex structures can be described in many ways, and they enable us to do complex analysis, geometry, and algebra on the surface X . In particular, they enable us to define HOLOMORPHIC [I.3 §5.6] (complex-analytic) and MEROMORPHIC FUNCTIONS [V.34] on open subsets of X . To be precise, X is a two-dimensional manifold, but the charts are thought of as open subsets of \mathbb{C} rather than of \mathbb{R} , and the maps that glue them together are required to be holomorphic. An equivalent notion is that of a *conformal structure* on X , which is the structure needed to make it possible to define angles between curves in X . Yet another important equivalent notion is that of *algebraic structure* on X , making X into a *complex-algebraic curve* (leading to the persistent confusion in terminology: a Riemann surface is two dimensional, and therefore a surface, from the point of view of topology or the real numbers, but one dimensional, and therefore a curve, from the point of view of complex analysis and algebra). An algebraic structure is what allows us to speak of polynomial, rational, or algebraic functions on X , and is usually specified by realizing X as the set of solutions to polynomial equations in complex PROJECTIVE SPACE [III.74] \mathbb{CP}^2 (or \mathbb{CP}^n).

In order to speak of a classification problem, let alone a moduli space, for Riemann surfaces we must next specify when we regard two Riemann surfaces as equivalent. (We postpone the discussion of the final ingredient, the notion of families of Riemann surfaces, to section 2.2.) To do this, we must give a notion of *isomorphism* between Riemann surfaces: when should two Riemann surfaces X and Y be “identified,” or thought of as giving two equivalent realizations of the same underlying object of our classification? This issue was hidden in our toy example of classifying lines in the plane: there we simply identified two lines if and only if they were *equal* as lines in the plane. This naive option is not available to us with the more abstractly defined Riemann surfaces. If we considered Riemann

PUP: ‘identified’ is an essential piece of jargon and is being used correctly in that sense here. Addition of quote marks hopefully mitigates the problem?

surfaces realized concretely as subsets of some larger space—for example, as solution sets to algebraic equations in complex projective space—we could similarly choose to identify surfaces only if they were equal as subsets. However, this is too fine a classification for most applications: what we care about is the *intrinsic geometry* of Riemann surfaces, and not incidental features that result from the particular way we choose to realize them.

At the other extreme, we might choose to ignore the extra geometric structure that makes a surface into a Riemann surface. That is, we could identify two Riemann surfaces X and Y if they are topologically equivalent, or homeomorphic (the “coffee mug is a doughnut” perspective). The classification of compact Riemann surfaces up to topological equivalence is captured by a single positive integer, the genus g (“number of holes”) of the surface. Any surface of genus zero is homeomorphic to the Riemann sphere $\mathbb{CP}^1 \simeq S^2$, any surface of genus 1 is homeomorphic to a torus $S^1 \times S^1$, and so on. Thus, in this case there is no issue of “modulation”—the classification is solved by giving a list of possible values of a single discrete invariant.

However, if we are interested in Riemann surfaces *as Riemann surfaces* rather than simply as topological manifolds, then this classification is too crude: it completely ignores the complex structure. We would now like to refine our classification to remedy this defect. To this end, we say that two Riemann surfaces X and Y are (conformally, or holomorphically) *equivalent* if there is a topological equivalence between them that preserves the geometry, i.e., a homeomorphism that preserves the angles between curves, or takes holomorphic functions to holomorphic functions, or takes rational functions to rational functions. (These conditions are all equivalent.) Note that we still have at our disposal our discrete invariant—the genus of a surface. However, as we shall see, this invariant is not fine enough to distinguish between all inequivalent Riemann surfaces. In fact, it is possible to have families of inequivalent Riemann surfaces that are parametrized by *continuous* parameters (but we cannot make proper sense of this idea until we have said precisely what is meant by a family of Riemann surfaces). Thus, the next step is to fix our discrete invariant and to try to classify all the different isomorphism classes of Riemann surfaces with the same genus by assembling them in a natural geometric fashion.

An important step toward this classification is the UNIFORMIZATION THEOREM [V.37]. This states that any simply connected Riemann surface is holomorphically

isomorphic to one of the following three: the Riemann sphere \mathbb{CP}^1 , the complex plane \mathbb{C} , or the upper half-plane \mathbb{H} (equivalently, the unit disk D). Since the UNIVERSAL COVERING SPACE [III.95] of any Riemann surface is a simply connected Riemann surface, the uniformization theorem provides an approach to classifying arbitrary Riemann surfaces. For instance, any COMPACT [III.9] Riemann surface of genus zero is simply connected, and in fact homeomorphic to the Riemann sphere, so the uniformization theorem already solves our classification problem in genus zero: up to equivalence, \mathbb{CP}^1 is the *only* Riemann surface of genus zero, and so in this case the topological and conformal classifications agree.

2.1 Moduli of Elliptic Curves

Next, we consider Riemann surfaces whose universal cover is \mathbb{C} , which is the same as saying that they are quotients of \mathbb{C} . For example, we can look at a quotient of \mathbb{C} by \mathbb{Z} , which means that we regard two complex numbers z and w as equivalent if $z - w$ is an integer. This has the effect of “wrapping \mathbb{C} around” into a cylinder. Cylinders are not compact, but to get a compact surface we could take a quotient by \mathbb{Z}^2 instead: that is, we could regard z and w as equivalent if their difference is of the form $a + bi$, where a and b are both integers. Now \mathbb{C} is wrapped around in two directions and the result is a torus with a complex (or, equivalently, conformal or algebraic) structure. This is a compact Riemann surface of genus 1. More generally, we can replace \mathbb{Z}^2 by any lattice L , regarding z and w as equivalent if $z - w$ belongs to L . (A *lattice* L in \mathbb{C} is an additive subgroup of \mathbb{C} with two properties. First, it is not contained in any line. Second, it is *discrete*, which means that there is a constant $d > 0$ such that the distance between any two points in L is at least d . Lattices are also discussed in THE GENERAL GOALS OF MATHEMATICAL RESEARCH [I.4 §4]. A *basis* for a lattice L is a pair of complex numbers u and v belonging to L such that every z in L can be written in the form $au + bv$ with a and b integers. Such a basis will not be unique: for example, if $L = \mathbb{Z} \oplus \mathbb{Z}$, then the obvious basis is $u = 1$ and $v = i$, but $u = 1$ and $v = 1 + i$ would do just as well.) If we take a quotient of \mathbb{C} by a lattice, then we again obtain a torus with complex structure. It turns out that any compact Riemann surface of genus 1 can be produced in this way.

From a topological point of view, any two tori are the same, but once we consider the complex structure we

start to find that different choices of lattice may lead to different Riemann surfaces. Certain changes to L do *not* have an effect: for example, if we multiply a lattice L by some nonzero complex number λ , then the quotient surface \mathbb{C}/L will not be affected. That is, \mathbb{C}/L is naturally isomorphic to $\mathbb{C}/\lambda L$. Therefore, we need only worry about the difference between lattices when one is not a multiple of the other. Geometrically, this says that one cannot be obtained from the other by a combination of rotation and dilation.

Notice that by taking the quotient \mathbb{C}/L we obtain not just a “naked” Riemann surface, but one equipped with an “origin,” that is, a distinguished point $e \in E$, which is the image of the origin $0 \in \mathbb{C}$. In other words, we obtain an *elliptic curve*:

Definition. An elliptic curve (over \mathbb{C}) is a Riemann surface E of genus 1, equipped with a marked point $e \in E$. Elliptic curves, up to isomorphism, are in bijection with lattices $L \subset \mathbb{C}$ up to rotation.

Remark. In fact, since $L \subset \mathbb{C}$ is a *subgroup* of the Abelian group \mathbb{C} , the elliptic curve $E = \mathbb{C}/L$ is naturally an Abelian group, with e as its identity element. This is an important motivation for keeping e as part of the data that defines an elliptic curve. A more subtle reason for remembering the location of e when we speak of E is that it helps us to define E more uniquely. This is useful, because any surface E of genus 1 has lots of symmetries, or AUTOMORPHISMS [I.3 §4.1]: there is always a holomorphic automorphism of E taking any point x to any other given point y . (If we think of E as a group, these are achieved by translations.) Thus, if someone hands us another genus-1 surface E' , there may be no way to identify E with E' , or there may be infinitely many ways: we can always compose a given isomorphism between them with a self-symmetry of E . As we will discuss later, automorphisms haunt almost every moduli problem, and are crucial when we consider the behavior of families. It is usually convenient to “rigidify” the situation somewhat, so that the possible isomorphisms between different objects are less “floppy” and more uniquely determined. In the case of elliptic curves, distinguishing the point e achieves this by reducing the symmetry of E . Once we do that, there is usually at most one way to identify two elliptic curves (one way, that is, that takes origin to origin).

We see that Riemann surfaces of genus 1 (with the choice of a marked point) can be described by concrete “linear algebra data”: a lattice $L \subset \mathbb{C}$, or rather the equivalence class consisting of all nonzero scalar multiples

λL of L . This is the ideal setting to study a classification, or moduli, problem. The next step is to find an explicit parametrization of the collection of all lattices, up to multiplication, and to decide in what sense we have obtained a geometric solution to the classification problem.

In order to parametrize the collection of lattices, we follow a procedure used for all moduli problems: first parametrize lattices together with the choice of some additional structure, and then see what happens when we forget this choice. For every lattice L we choose a basis $\omega_1, \omega_2 \in L$: that is, we represent L as the set of all integer combinations $a\omega_1 + b\omega_2$. We do this in an *oriented* fashion: we require that the *fundamental parallelogram* spanned by ω_1 and ω_2 is positively oriented. (That is, the numbers $0, \omega_1, \omega_1 + \omega_2$, and ω_2 list the vertices of the parallelogram in a counterclockwise order. From the geometric point of view of the elliptic curve E , L is the FUNDAMENTAL GROUP [IV.6 §2] of E , and the orientation condition says that we generate L by two loops, or “meridians,” $A = \omega_1, B = \omega_2$, which are oriented, in that their oriented intersection number $A \cap B$ is equal to $+1$ rather than -1 .) Since we are interested in lattices only up to multiplication, we can multiply L by a complex number so as to turn ω_1 into 1 and hence ω_2 into $\omega = \omega_2/\omega_1$. The orientation condition now says that ω is in the upper half-plane \mathbb{H} : i.e., its imaginary part is positive, $\text{Im } \omega > 0$. Conversely, any complex number $\omega \in \mathbb{H}$ in the upper half-plane determines a unique oriented lattice $L = \mathbb{Z}1 \oplus \mathbb{Z}\omega$ (that is, the set of all integer combinations $a + b\omega$ of 1 and ω) and no two of these lattices are related by a rotation.

What does this tell us about elliptic curves? We saw earlier that an elliptic curve is defined by a lattice L and an identity e . Now we have seen that if we give L some extra structure, namely an oriented basis, then we can parametrize it by a complex number $\omega \in \mathbb{H}$. This makes precise for us the “additional structure” that we want to place on elliptic curves. We say that a *marked* elliptic curve is an elliptic curve E, e together with the choice of an oriented basis ω_1, ω_2 for the associated lattice (fundamental group) L of E . The point is that any lattice has infinitely many different bases, which lead to many automorphisms of E . By “marking” one of these bases, we stop them being automorphisms.

2.2 Families and Teichmüller Spaces

With our new definition, we can summarize the earlier discussion by saying that marked elliptic curves are in

PUP: Tim would like to keep this sentence as it is. OK?

bijection with points $\omega \in \mathbb{H}$ of the upper half-plane. The upper half-plane is, however, much more than just a set of points: it carries a host of geometric structures, in particular a topology and a complex structure. In what sense do these structures reflect geometric properties of marked elliptic curves? In other words, in what sense is the complex manifold \mathbb{H} , known in this context as the *Teichmüller space* $\mathcal{T}_{1,1}$ of genus-1 Riemann surfaces with one marked point, a geometric solution to the problem of classifying marked elliptic curves?

In order to answer this question, we need the notion of a continuous family of Riemann surfaces, and also the notion of a complex-analytic family. A *continuous family of Riemann surfaces* parametrized by a topological space S , such as the circle S^1 , for example, is a “continuously varying” assignment of a Riemann surface X_s to every point s of S . In our example of the moduli of lines in the plane, a continuous family of lines was characterized by the property that the angles between the lines and the x -axis or y -axis defined continuous functions of the parameters. Geometrically defined collections of lines, such as those produced by a curve C in the plane, then gave rise to continuous families. More abstractly, a continuous family of lines defined a line *bundle* over the parameter space. A good criterion for a family of Riemann surfaces is likewise that any “reasonably defined” geometric quantity that we can calculate for every Riemann surface should vary continuously in the family. For example, a classical construction of Riemann surfaces of genus g comes from taking $4g$ -gons and gluing opposite sides together. The resulting Riemann surface is fully determined by the edge-lengths and angles of the polygon. Therefore, a continuous family of Riemann surfaces described in this fashion should be precisely a family such that the edge-lengths and angles give *continuous* functions of the parameter set.

In more abstract topological terms, if we have a collection $\{X_s, s \in S\}$ of Riemann surfaces depending on points in a space S and we wish to make it into a continuous family, then we should give the union $\bigcup_{s \in S} X_s$ itself the structure of a topological space X , which should simultaneously extend the topology on each individual X_s . The result is called a *Riemann surface bundle*. Associated with X is the map that takes each point x to the particular s for which x belongs to X_s . We should demand that this map is continuous, and perhaps more (it could be a fibration, or fiber bundle). This definition has the advantage of great flexibility. For example, if S is a complex manifold, then in just

the same way we can speak of a *complex-analytic family of Riemann surfaces* $\{X_s, s \in S\}$ parametrized by S : now we ask for the union of the X_s to carry not just a topology but a complex structure (i.e., it should form a complex manifold), extending the complex structure on the fibers and mapping holomorphically to the parameter set. The same holds with “complex-analytic” replaced by “algebraic.” These abstract definitions have the property that if our Riemann surfaces are described in a concrete way—cut out by equations, glued from coordinate patches, etc.—then the coefficients of our equations or gluing data will vary as complex-analytic functions in our family precisely when the family is complex analytic (and likewise for continuous or algebraic families).

As a reality check, note that a (continuous, analytic, or other) family of Riemann surfaces parametrized by a single point $s = S$ is indeed just a single Riemann surface X_s . Just as in this simple case we wish to consider Riemann surfaces only up to equivalence, so there is a notion of equivalence or isomorphism of two analytic families $\{X_s\}$ and $\{X'_s\}$ parametrized by the same space S . We simply regard the families as equivalent if the surfaces X_s and X'_s are isomorphic for every s , and if the isomorphism depends analytically on s .

Armed with the notion of family, we can now formulate the characteristic property that the upper half-plane possesses when we think of it as the moduli space of marked elliptic curves. We define a continuous or analytic family of marked elliptic curves to be a family where the underlying genus-1 surfaces vary continuously or analytically, while the choice of basepoint $e_s \in E_s$ and the basis of the lattice L_s vary continuously.

The upper half-plane \mathbb{H} plays a role for marked elliptic curves that is similar to the role played by \mathbb{RP}^1 for lines in the plane. The following theorem makes this statement precise.

Theorem. *For any topological space S , there is a one-to-one correspondence between continuous maps from S to \mathbb{H} and isomorphism classes of continuous families of marked elliptic curves parametrized by S . Similarly, there is a one-to-one correspondence between analytic maps from any complex manifold S to \mathbb{H} and isomorphism classes of analytic families of marked elliptic curves parametrized by S .*

If we apply the theorem in the case where S is a single point, it simply tells us that the points of \mathbb{H} are in bijection with the isomorphism classes of marked elliptic curves, as we already knew. However, it contains more

information: it says that \mathbb{H} , with its topology and complex structure, *embodies the structure* of marked elliptic curves and the ways in which they can modulate. At the other extreme, we could take $S = \mathbb{H}$ itself, mapping S to \mathbb{H} by the identity map. This expresses the fact that \mathbb{H} itself carries a family of marked elliptic curves, i.e., the collection of Riemann surfaces defined by $\omega \in \mathbb{H}$ fit together into a complex manifold fibering over \mathbb{H} with elliptic curve fibers. This family is called the *universal family*, since by the theorem any family is “deduced” (or pulled back) from this one universal example.

2.3 From Teichmüller Spaces to Moduli Spaces

We have arrived at a complete and satisfying picture for the classification of elliptic curves when we choose in addition a marking (that is, an oriented basis of the associated lattice $L = \pi_1(E)$). What can we say about elliptic curves themselves, without the choice of marking? We somehow need to “forget” the marking, by regarding two points of \mathbb{H} as equivalent if they correspond to two different markings of the same elliptic curve.

Now, given any two bases of the group (or lattice) $\mathbb{Z} \oplus \mathbb{Z}$, there is an invertible 2×2 matrix with integer entries that takes one basis to the other. If the two bases are *oriented*, then this matrix will have determinant 1, which means that it is an element

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$$

of the group of invertible unimodular matrices over \mathbb{Z} . Similarly, given any two oriented bases (ω_1, ω_2) and (ω'_1, ω'_2) of a lattice L , which can be thought of as oriented identifications of L with $\mathbb{Z} \oplus \mathbb{Z}$, there is a matrix $A \in \mathrm{SL}_2(\mathbb{Z})$ such that $\omega'_1 = a\omega_1 + b\omega_2$ and $\omega'_2 = c\omega_1 + d\omega_2$. If we now consider the normalized bases $(1, \omega)$ and $(1, \omega')$, where $\omega = \omega_1/\omega_2$ and $\omega' = \omega'_1/\omega'_2$, then we obtain a transformation of the upper half-plane. It is given by the formula

$$\omega' = \frac{a\omega + b}{c\omega + d}.$$

That is, the group $\mathrm{SL}_2(\mathbb{Z})$ is acting on the upper half-plane by linear fractional (or Möbius) transformations with integer coefficients, and two points in the upper half-plane correspond to the same elliptic curve if one can be turned into the other by means of such a transformation. If this is the case, then we should regard the two points as equivalent: that is how we formalize the idea of “forgetting” the marking. Note also that the scalar matrix $-\mathrm{Id}$ in $\mathrm{SL}_2(\mathbb{Z})$, which negates both ω_1 and

ω_2 , acts trivially on the upper half-plane, so that we in fact get an action of $\mathrm{PSL}_2(\mathbb{Z}) = \mathrm{SL}_2(\mathbb{Z}) / \{\pm \mathrm{Id}\}$ on \mathbb{H} .

So we come to the conclusion that *elliptic curves (up to isomorphism) are in bijection with orbits of $\mathrm{PSL}_2(\mathbb{Z})$ on the upper half-plane, or equivalently with points of the quotient space $\mathbb{H} / \mathrm{PSL}_2(\mathbb{Z})$* . This quotient space has a natural quotient topology, and in fact can be given a complex-analytic structure, which, it turns out, identifies it with the complex plane \mathbb{C} itself. To see this one uses the classical MODULAR FUNCTION [IV.1 §8] $j(z)$, a complex-analytic function on \mathbb{H} which is invariant under the modular group $\mathrm{PSL}_2(\mathbb{Z})$ and which therefore defines a natural coordinate $\mathbb{H} / \mathrm{PSL}_2(\mathbb{Z}) \rightarrow \mathbb{C}$.

It appears that we have solved the moduli problem for elliptic curves: we have a topological, and even complex-analytic, space $\mathfrak{M}_{1,1} = \mathbb{H} / \mathrm{PSL}_2(\mathbb{Z})$ whose points are in one-to-one correspondence with isomorphism classes of elliptic curves. This already qualifies $\mathfrak{M}_{1,1}$ as the *coarse moduli space* for elliptic curves, which means it is as good a moduli space as we can hope for. However, $\mathfrak{M}_{1,1}$ fails an important test for a moduli space that $\mathcal{T}_{1,1}$ passed (as we saw in section 2.2): it is *not* true, even for the circle $S = S^1$, that every continuous family of elliptic curves over S corresponds to a map from S to $\mathfrak{M}_{1,1}$.

The reason for this failure is the problem of automorphisms. These are equivalences from E to itself: that is, complex-analytic maps from E to E that preserve the basepoint e . Equivalently, they are given by complex-analytic self-maps of \mathbb{C} that preserve 0 and the lattice L . Such a map must be a rotation: that is, multiplication by some complex number λ of modulus 1. It is easy to check that for most lattices L in the plane, the only rotation that sends L to itself is multiplication by $\lambda = -1$. Note that this is the same -1 that we quotiented out by to pass from $\mathrm{SL}_2(\mathbb{Z})$ to $\mathrm{PSL}_2(\mathbb{Z})$. However, there are two special lattices that have greater symmetry. These are the *square lattice* $L = \mathbb{Z} \cdot 1 \oplus \mathbb{Z} \cdot i$, corresponding to the fourth root of unity i , and the *hexagonal lattice* $L = \mathbb{Z} \cdot 1 \oplus \mathbb{Z} \cdot e^{2\pi i/6}$, corresponding to a sixth root of unity. (Note that the hexagonal lattice is also represented by the point $\omega = e^{2\pi i/3}$.) The square lattice, which corresponds to the elliptic curve formed by gluing the opposite sides of a square, has as its symmetries the group $\mathbb{Z}/4\mathbb{Z}$ of rotational symmetries of the square. The hexagonal lattice, which corresponds to the elliptic curve formed by gluing the opposite sides of a regular hexagon, has as its symmetries the group $\mathbb{Z}/6\mathbb{Z}$ of rotational symmetries of a hexagon.

PUP: I can confirm that this sentence is correct as written.

We see that the number of automorphisms of an elliptic curve jumps discontinuously at the special points $\omega = i$ and $\omega = e^{2\pi i/6}$. This already suggests that something might be wrong with $\mathfrak{M}_{1,1}$ as a moduli space. Note that we avoided this problem with the moduli $\mathcal{T}_{1,1}$ of *marked* elliptic curves, since there are no automorphisms of an elliptic curve that also preserve the marking. Another place we might have observed this problem with $\mathfrak{M}_{1,1}$ is when we passed to the quotient $\mathbb{H}/\mathrm{PSL}_2(\mathbb{Z})$. We avoided the automorphism $\lambda = -1$ by quotienting by $\mathrm{PSL}_2(\mathbb{Z})$ rather than $\mathrm{SL}_2(\mathbb{Z})$. However, the two special points i and $e^{2\pi i/6}$ are preserved by integer Möbius transformations of \mathbb{H} other than the identity, and they are the only points with that property. This means that the quotient $\mathbb{H}/\mathrm{PSL}_2(\mathbb{Z})$ naturally comes with conical singularities at the points corresponding to these two orbits: one looks like a cone with angle π , and the other like a cone with angle $\frac{2}{3}\pi$. (To see why this is plausible, imagine the following simpler instance of the same phenomenon. If for every complex number z you identify z with $-z$, then the result is to wrap the complex plane around into a cone with a singularity at 0. The reason 0 is singled out is that it is preserved by the transformation $z \mapsto -z$. Here the angle would be π because the identification of points is two-to-one away from the singularity and π is half of 2π .) It is possible to massage these singularities away using the j -function, but they are indicating a basic difficulty.

So why do automorphisms form an obstacle to the existence of “good” moduli spaces? We can demonstrate the difficulty by considering an interesting continuous family of marked elliptic curves parametrized by the circle $S = S^1$. Let $E(i)$ be the “square” elliptic curve that we considered earlier, based on the lattice of integer combinations of 1 and i . Next, for every t between 0 and 1, let E_t be a copy of $E(i)$. Thus, we have taken the constant, or “trivial,” family of elliptic curves over the closed unit interval $[0, 1]$, where every curve in the family is $E(i)$. Now we identify the elliptic curves at the two ends of this family, not in the obvious way, but by using the automorphism given by a 90° rotation, or multiplication by i . This means that we are looking at the family of elliptic curves over the circle where each member of the family is a copy of the elliptic curve $E(i)$, but these copies twist by 90° as we go around the circle.

It is easy to see that there is no way to capture this family of elliptic curves by means of a map from S^1 to the space $\mathfrak{M}_{1,1}$. Since all of the members of the family are isomorphic, each point of the circle should map to the same point in $\mathfrak{M}_{1,1}$ (the equivalence class of i in

\mathbb{H}). But the constant map $S^1 \rightarrow \{i\} \in \mathfrak{M}_{1,1}$ classifies the *trivial* family $S^1 \times E_i$ of elliptic curves over S^1 , that is, the family where every curve is equal to $E(i)$ but the curves *do not* twist as we go around! Thus, there are more families of elliptic curves than there are maps to $\mathfrak{M}_{1,1}$; the quotient space $\mathbb{H}/\mathrm{PSL}_2(\mathbb{Z})$ cannot handle the complications caused by automorphisms. A variant of this construction applies to complex-analytic families with S^1 replaced by \mathbb{C}^\times . This is a very general phenomenon in moduli problems: whenever objects have nontrivial automorphisms, we can imitate the construction above to get nontrivial families over an interesting parameter set, all of whose members are the same. As a result, they cannot be classified by a map to the set of all isomorphism classes.

What do we do about this problem? One approach is to resign ourselves to having coarse moduli spaces, which have the right points and right geometry but do not quite classify arbitrary families. Another approach is the one that leads to $\mathcal{T}_{1,1}$: we can fix markings of one kind or another, which “kill” all automorphisms. In other words, we choose enough extra structure on our objects so that there do not remain any (nontrivial) automorphisms that preserve all this decoration. In fact, one can be far more economical than picking a basis of the lattice L and obtaining the infinite covering $\mathcal{T}_{1,1}$ of $\mathfrak{M}_{1,1}$: one can fix a basis of L only up to some congruence (for example, of $L/2L$). Finally, we can simply learn to come to terms with the automorphisms, keeping them as part of the data, resulting in “spaces” where points have internal symmetries. This is the notion of an ORBIFOLD [IV.4 §7], or STACK [IV.4 §7], which is flexible enough to deal with essentially all moduli problems.

3 Higher-Genus Moduli Spaces and Teichmüller Spaces

We would now like to generalize as much as possible of the picture of elliptic curves and their moduli to higher-genus Riemann surfaces. For each g we would like to define a space \mathfrak{M}_g , called the *moduli space of curves of genus g* , that classifies compact Riemann surfaces of genus g and tells us how they modulate. Thus, the points of \mathfrak{M}_g should correspond to our objects, compact Riemann surfaces of genus g , or, to be more accurate, equivalence classes of such surfaces, where two surfaces are considered to be equivalent if there is a complex-analytic isomorphism between them. In addition, we would like \mathfrak{M}_g to do the best

PUP: again, all fine here.

it can to embody the structure of continuous families of genus- g surfaces. Likewise, there are spaces $\mathcal{M}_{g,n}$ parametrizing “ n -punctured” Riemann surfaces of genus g . This means we consider not “bare” Riemann surfaces, but Riemann surfaces together with a “decoration” or “marking” by n distinct labeled points (punctures). Two of these are considered to be equivalent if there is a complex-analytic isomorphism between them that takes punctures to punctures and preserves labels. Since there are Riemann surfaces with automorphisms, we do not expect \mathcal{M}_g to be able to classify all families of Riemann surfaces: that is, we will expect examples similar to the twisted square-lattice construction discussed earlier. However, if we consider Riemann surfaces with enough extra markings, then we will be able to obtain a moduli space in the strongest sense. One way to choose such markings is to consider $\mathcal{M}_{g,n}$ with n large enough (for fixed g). Another approach will be to mark generators of the fundamental group, leading to the Teichmüller spaces \mathcal{T}_g and $\mathcal{T}_{g,n}$. We now outline this process.

To construct the space \mathcal{M}_g , we return to the uniformization theorem. Any compact surface X of genus $g > 1$ has as its universal cover the upper half-plane \mathbb{H} , so it is represented as a quotient $X = \mathbb{H}/\Gamma$, where Γ is a representation of the fundamental group of X as a subgroup of conformal self-maps of \mathbb{H} . The group of all conformal automorphisms of \mathbb{H} is $\mathrm{PSL}_2(\mathbb{R})$, the group of linear fractional transformations with real coefficients. The fundamental groups of all compact genus- g Riemann surfaces are isomorphic to a fixed abstract group Γ_g , with $2g$ generators A_i, B_i ($i = 1, \dots, g$) and one relation: that the product of all commutators $A_i B_i A_i^{-1} B_i^{-1}$ is the identity. A subgroup $\Gamma \subset \mathrm{PSL}_2(\mathbb{R})$ that acts on \mathbb{H} in such a way that the quotient \mathbb{H}/Γ is a Riemann surface (technically, the action should have no fixed points and should be properly discontinuous) is known as a FUCHSIAN GROUP [III.28]. Thus, the analogue of the representation of elliptic curves by lattices $L \simeq \mathbb{Z} \oplus \mathbb{Z}$ in the plane is the representation of higher-genus Riemann surfaces as \mathbb{H}/Γ , where Γ is a Fuchsian group.

The Teichmüller space \mathcal{T}_g of genus- g Riemann surfaces is the space that solves the moduli problem for genus- g surfaces when they come with a marking of their fundamental group. This means that our objects are genus- g surfaces X plus a set of generators A_i, B_i of $\pi_1(X)$, which give an isomorphism between $\pi_1(X)$ and Γ_g , up to conjugation.¹ Our equivalences

are complex-analytic maps that preserve the markings. Finally, our continuous (respectively, complex-analytic) families are continuous (complex-analytic) families of Riemann surfaces with continuously varying markings of the fundamental group. In other words, we are asserting the existence of a topological space/complex manifold \mathcal{T}_g , with a complex-analytic family of marked Riemann surfaces over it, and the following strong property.

The characteristic property of \mathcal{T}_g . *For any topological space (respectively, complex manifold) S , there is a bijection between continuous maps (respectively, holomorphic maps) $S \rightarrow \mathcal{T}_g$ and isomorphism classes of continuous (respectively, complex-analytic) families of marked genus- g surfaces parametrized by S .*

3.1 Digression: “Abstract Nonsense”

It is interesting to note that, while we have yet to see why such a space exists, it follows from general, nongeometric principles—CATEGORY THEORY [III.8] or “abstract nonsense”—that it is completely and uniquely determined, both as a topological space and as a complex manifold, by this characteristic property. In a very abstract way, every topological space M can be uniquely reconstructed from its set of points, the set of paths between these points, the set of surfaces spanning these paths, and so on. To put it differently, we can think of M as a “machine” that assigns to any topological space S the set of continuous maps from S to M . This machine is known as the “functor of points of M .” Similarly, a complex manifold M provides a machine that assigns to any other complex manifold S the set of complex-analytic maps from S to M . A curious discovery of category theory (the *Yoneda lemma*) is that for very general reasons (having nothing to do with geometry), these machines (or functors) uniquely determine M as a space, or a complex manifold.

Any moduli problem in the sense we have described (giving objects, equivalences, and families) also gives such a machine, where to S we assign the set of all families over S , up to isomorphism. So *just by setting up the moduli problem* we have already uniquely determined the topology and complex structure on Teichmüller space. The interesting part then is to know whether or not there *actually exists* a space giving rise to the same

1. Note that while the fundamental group of X depends on the choice of a basepoint, $\pi_1(X, x)$ and $\pi_1(X, y)$ may be identified by

choosing a path from x to y , and the different choices are related by conjugation by a loop. Thus, if we are willing to identify sets of generators A_i, B_i when they differ only by a conjugation, then we can ignore the choice of a basepoint.

machine we have constructed, whether we can construct it explicitly, and whether we can use its geometry to learn interesting facts about Riemann surfaces.

3.2 Moduli Spaces and Representations

Coming back to earth, we discover that we have a fairly concrete model of Teichmüller space at our disposal. Once we have fixed the marking $\pi_1(X) \simeq \Gamma_g$, we are simply looking at all ways to represent Γ_g as a Fuchsian subgroup of $\mathrm{PSL}_2(\mathbb{R})$. Ignoring the Fuchsian condition for a moment, this means finding $2g$ real matrices (up to $\pm \mathrm{Id}$) $A_i, B_i \in \mathrm{PSL}_2(\mathbb{R})$ satisfying the commutator relation of Γ_g . This gives an explicit set of (algebraic!) equations for the entries of the $2g$ matrices, which determine the space of all representations $\Gamma_g \rightarrow \mathrm{PSL}_2(\mathbb{R})$. We must now quotient out by the action of $\mathrm{PSL}_2(\mathbb{R})$ that simultaneously conjugates all $2g$ matrices to obtain the *representation variety* $\mathrm{Rep}(\Gamma_g, \mathrm{PSL}_2(\mathbb{R}))$. This is analogous to considering lattices in \mathbb{C} up to rotation, and is motivated by the fact that the quotients of \mathbb{H} by two conjugate subgroups of $\mathrm{PSL}_2(\mathbb{R})$ will be isomorphic.

Once we have described the space of all representations of Γ_g into $\mathrm{PSL}_2(\mathbb{R})$, we can then single out Teichmüller space as the subset of the representation variety that consists of Fuchsian representations of Γ_g into $\mathrm{PSL}_2(\mathbb{R})$. Luckily this subset is *open* in the representation variety, which gives a nice realization of \mathcal{T}_g as a topological space—in fact, \mathcal{T}_g is homeomorphic to \mathbb{R}^{6g-6} . (This can be seen very explicitly in terms of the *Fenchel-Nielsen* coordinates, which parametrize a surface in \mathcal{T}_g via a cut-and-paste procedure involving $3g - 3$ lengths and $3g - 3$ angles.) We may now try to “forget” the marking $\pi_1(X) \cong \Gamma_g$, to obtain the moduli space \mathfrak{M}_g of unmarked Riemann surfaces. In other words, we would like to take \mathcal{T}_g and identify any two points that represent the same underlying Riemann surface with different markings. This identification is achieved by the action of a group, the *genus- g mapping class group* MCG_g or *Teichmüller modular group*, on \mathcal{T}_g , which generalizes the modular group $\mathrm{PSL}_2(\mathbb{Z})$ that acts on $\mathbb{H} = \mathcal{T}_{1,1}$. (The mapping class group is defined as the group of all self-diffeomorphisms of a genus- g surface—remember that all such surfaces are topologically the same—modulo those diffeomorphisms that act trivially on the fundamental group.) As in the case of elliptic curves, Riemann surfaces with automorphisms correspond to points in \mathcal{T}_g fixed by some subgroup of MCG_g , and give rise to singular points in the quotient $\mathfrak{M}_g = \mathcal{T}_g / \mathrm{MCG}_g$.

Representation varieties, or moduli spaces of representations, are an important and concrete class of moduli spaces that arise throughout geometry, topology, and number theory. Given any (discrete) group Γ , we ask (for example) for a space that parametrizes homomorphisms of Γ into the group of $n \times n$ matrices. The notion of equivalence is given by conjugation by GL_n , and that of families by continuous (or analytic, or algebraic, etc.) families of matrices. This problem is interesting even when the group Γ is \mathbb{Z} . Then we are simply considering invertible $n \times n$ matrices (the image of $1 \in \mathbb{Z}$) up to conjugacy. It turns out that there is no moduli space for this problem, even in the coarse sense, unless we consider only “nice enough” matrices: for example, matrices that consist of only a single Jordan block. This is a good example of a ubiquitous phenomenon in moduli problems: one is often forced to throw out some “bad” (unstable) objects in order to have any chance of obtaining a moduli space. (See the paper by Mumford and Suominen (1972) for a detailed discussion.)

3.3 Moduli Spaces and Jacobians

The upper half-plane $\mathbb{H} = \mathcal{T}_{1,1}$, together with the action of $\mathrm{PSL}_2(\mathbb{Z})$, gives an appealingly complete picture of the moduli problem for elliptic curves and its geometry. The same cannot be said, unfortunately, for the picture of \mathcal{T}_g as an open subset of the representation variety. In particular, the representation variety does not even carry a natural complex structure, so we cannot see from this description the geometry of \mathcal{T}_g as a complex manifold. This failure reflects some of the ways in which the study of moduli spaces is more complicated for genus greater than 1. In particular, the moduli spaces of higher-genus surfaces are not described purely by linear algebra plus data about orientation, as is the case in genus 1.

Part of the blame for this complexity lies with the fact that the fundamental group $\Gamma_g \simeq \pi_1(X)$ ($g > 1$) is no longer Abelian, and in particular it is no longer equal to the first homology group $H_1(X, \mathbb{Z})$. A related problem is that X is no longer a group. A beautiful solution to this problem is given by the construction of the Jacobian $\mathrm{Jac}(X)$, which shares with elliptic curves the properties of being a torus (homeomorphic to $(S^1)^{2g}$), an Abelian group, and a complex (in fact complex-algebraic) manifold. (The Jacobian of an elliptic curve is the elliptic curve itself.) The Jacobian captures the “Abelian” or “linear” aspects of the geometry of X . There is a moduli space \mathcal{A}_g for such complex-algebraic tori (known as

Abelian varieties), which does share all of the nice properties and linear algebraic description of the moduli of elliptic curves $\mathcal{M}_{1,1} = \mathcal{A}_1$. The good news—the Torelli theorem—is that by assigning to each Riemann surface X its Jacobian we embed \mathcal{M}_g as a closed, complex-analytic subset of \mathcal{A}_g . The *interesting* news—the Schottky problem—is that the image is quite complicated to characterize intrinsically. In fact, solutions to this problem have come from as far afield as the study of nonlinear partial differential equations!

3.4 Further Directions

In this section we give hints at some interesting questions about, and applications of, moduli spaces.

Deformations and degenerations. Two of the main topics in moduli spaces ask which objects are very near to a given one, and what lies very far away. Deformation theory is the calculus of moduli spaces: it describes their infinitesimal structure. In other words, given an object, deformation theory is concerned with describing all its small perturbations (see Mazur (2004) for a beautiful discussion of this). At the other extreme, we can ask what happens when our objects degenerate? Most moduli spaces, for example the moduli of curves, are not compact, so there are families “going off to infinity.” It is important to find “meaningful” compactifications of moduli spaces, which classify the possible degenerations of our objects. Another advantage of compactifying moduli spaces is that we can then calculate integrals over the completed space. This is crucial for the next item.

Invariants from moduli spaces. An important application of moduli spaces in geometry and topology is inspired by quantum field theory, where a particle, rather than following the “best” classical path between two points, follows all paths with varying probabilities (see MIRROR SYMMETRY [IV.16 §2.2.4]). Classically, one calculates many topological invariants by picking a geometric structure (such as a metric) on a space, calculating some quantity using this structure, and finally proving that the result of the calculation did not depend on the structure we chose. The new alternative is to look at *all* such geometric structures, and integrate some quantity over the space of all choices. The result, if we can show convergence, will manifestly not depend on any choices. String theory has given rise to many important applications of this idea, in particular by giving a rich structure to the collection of integrals

obtained in this way. Donaldson and Seiberg–Witten theories use this philosophy to give topological invariants of four-manifolds. Gromov–Witten theory applies it to the topology of SYMPLECTIC MANIFOLDS [III.90], and to counting problems in algebraic geometry, such as, How many rational plane curves of degree 5 pass through fourteen points in general position? (Answer: 87 304.)

Modular forms. One of the most profound ideas in mathematics, the Langlands program, relates number theory to function theory (harmonic analysis) on very special moduli spaces, generalizing the moduli space of elliptic curves. These moduli spaces (Shimura varieties) are expressible as quotients of symmetric spaces (such as \mathbb{H}) by arithmetic groups (such as $\mathrm{PSL}_2(\mathbb{Z})$). MODULAR FORMS [III.61] and automorphic forms are special functions on these moduli spaces, described by their interaction with the large symmetry groups of the spaces. This is an extremely exciting and active area of mathematics, which counts among its recent triumphs the proof of FERMAT’S LAST THEOREM [V.12] and the Shimura–Taniyama–Weil conjecture (Wiles, Taylor–Wiles, Breuil–Conrad–Diamond–Taylor).

Further Reading

For historical accounts and bibliographies on moduli spaces, the following articles are highly recommended.

A beautiful and accessible overview of moduli spaces, with an emphasis on the notion of deformations, is given by Mazur (2004). The articles by Hain (2000) and Looijenga (2000) give excellent introductions to the study of the moduli spaces of curves, perhaps the oldest and most important of all moduli problems. The article by Mumford and Suominen (1972) introduces the key ideas underlying the study of moduli spaces in algebraic geometry.

- Hain, R. 2000. Moduli of Riemann surfaces, transcendental aspects. In *School on Algebraic Geometry, Trieste, 1999*, pp. 293–353. ICTP Lecture Notes Series, no. 1. Trieste: The Abdus Salam International Centre for Theoretical Physics.
- Looijenga, E. 2000. A minicourse on moduli of curves. In *School on Algebraic Geometry, Trieste, 1999*, pp. 267–91. ICTP Lecture Notes Series, no. 1. Trieste: The Abdus Salam International Centre for Theoretical Physics.
- Mazur, B. 2004. Perturbations, deformations and variations (and “near-misses”) in geometry. Physics and number theory. *Bulletin of the American Mathematical Society* 41(3):307–36.

Mumford, D., and K. Suominen. 1972. Introduction to the theory of moduli. In *Algebraic Geometry, Oslo, 1970: Proceedings of the Fifth Nordic Summer School in Mathematics*, edited by F. Oort, pp. 171–222. Groningen: Wolters-Noordhoff.

IV.9 Representation Theory

Ian Grojnowski

1 Introduction

It is a fundamental theme in mathematics that many objects, both mathematical and physical, have symmetries. The goal of GROUP [I.3 §2.1] theory in general, and representation theory in particular, is to study these symmetries. The difference between representation theory and general group theory is that in representation theory one restricts one's attention to symmetries of VECTOR SPACES [I.3 §2.3]. I will attempt here to explain why this is sensible and how it influences our study of groups, causing us to focus on groups with certain nice structures involving *conjugacy classes*.

2 Why Vector Spaces?

The aim of representation theory is to understand how the *internal* structure of a group controls the way it acts *externally* as a collection of symmetries. In the other direction, it also studies what one can learn about a group's internal structure by regarding it as a group of symmetries.

We begin our discussion by making more precise what we mean by “acts as a collection of symmetries.” The idea we are trying to capture is that if we are given a group G and an object X , then we can associate with each element g of G some symmetry of X , which we call $\phi(g)$. For this to be sensible, we need the composition of symmetries to work properly: that is, $\phi(g)\phi(h)$ (the result of applying $\phi(h)$ and then $\phi(g)$) should be the same symmetry as $\phi(gh)$. If X is a set, then a symmetry of X is a particular kind of PERMUTATION [III.70] of its elements. Let us denote by $\text{Aut}(X)$ the group of *all* permutations of X . Then an *action* of G on X is defined to be a homomorphism from G to $\text{Aut}(X)$. If we are given such a homomorphism, then we say that G *acts* on X .

The image to have in mind is that G “does things” to X . This idea can often be expressed more conveniently and vividly by forgetting about ϕ in the notation: thus, instead of writing $\phi(g)(x)$ for the effect on x of the symmetry associated with g , we simply think of g itself

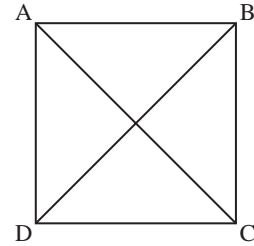


Figure 1 A square and its diagonals.

as a permutation and write gx . However, sometimes we do need to talk about ϕ as well: for instance, we might wish to compare two different actions of G on X .

Here is an example. Take as our object X a square in the plane, centered at the origin, and let its vertices be A, B, C , and D (see figure 1). A square has eight symmetries: four rotations by multiples of 90° and four reflections. Let G be the group consisting of these eight symmetries; this group is often called D_8 , or the *dihedral group* of order 8. By definition, G acts on the square. But it also acts on the set of *vertices* of the square: for instance, the action of the reflection through the y -axis is to switch A with B and C with D . It might seem as though we have done very little here. After all, we defined G as a group of symmetries so it does not take much effort to associate a symmetry with each element of G . However, we did not define G as a group of permutations of the set $\{A, B, C, D\}$, so we have at least done something.

To make this point clearer, let us look at some other sets on which G acts, which will include any set that we can build sufficiently naturally from the square. For instance, G acts not only on the set of vertices $\{A, B, C, D\}$, but on the set of edges $\{AB, BC, CD, DA\}$ and on the set of cross-diagonals $\{AC, BD\}$ as well. Notice in the latter case that some of the elements of G act in the same way: for example, a clockwise rotation through 90° interchanges the two diagonals, as does a counterclockwise rotation through 90° . If all the elements of G act differently, then the action is called *faithful*.

Notice that the operations on the square (“reflect through the y -axis,” “rotate through 90° ,” and so on) can be applied to the whole Cartesian plane \mathbb{R}^2 . Therefore, \mathbb{R}^2 is another (and much larger) set on which G acts. To call \mathbb{R}^2 a set, though, is to forget the very interesting fact that the elements in \mathbb{R}^2 can be added together and multiplied by real numbers: in other words, \mathbb{R}^2 is a *vector space*. Furthermore, the action

of G is well-behaved with respect to this extra structure. For instance, if g is one of our symmetries and v_1 and v_2 are two elements of \mathbb{R}^2 , then g applied to the sum $v_1 + v_2$ yields the sum $g(v_1) + g(v_2)$. Because of this, we say that G acts *linearly* on the vector space \mathbb{R}^2 . When V is a vector space, we denote by $\text{GL}(V)$ the set of invertible linear maps from V to V . If V is the vector space \mathbb{R}^n , this group is the familiar group $\text{GL}_n(\mathbb{R})$ of invertible $n \times n$ matrices with real entries; similarly, when $V = \mathbb{C}^n$ it is the group of invertible matrices with complex entries.

Definition. A *representation* of a group G on a vector space V is a homomorphism from G to $\text{GL}(V)$.

In other words, a group action is a way of regarding a group as a collection of permutations, while a representation is the special case where these permutations are invertible linear maps. One sometimes sees representations referred to, for emphasis, as *linear* representations. In the representation of D_8 on \mathbb{R}^2 that we described above, the homomorphism from G to $\text{GL}_2(\mathbb{R})$ took the symmetry “clockwise rotation through 90° ” to the matrix $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ and the symmetry “reflection through the y -axis” to the matrix $\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$.

Given one representation of G , we can produce others using natural constructions from linear algebra. For example, if ρ is the representation of G on \mathbb{R}^2 described above, then $\det \rho$ (see DETERMINANTS [III.15]) is a homomorphism from G to \mathbb{R}^* (the group of nonzero real numbers under multiplication), since

$$\det(\rho(gh)) = \det(\rho(g)\rho(h)) = \det(\rho(g)) \det(\rho(h)),$$

by the multiplicative property of determinants. This makes $\det \rho$ a one-dimensional representation, since each nonzero real number t can be thought of as the element “multiply by t ” of $\text{GL}_1(\mathbb{R})$. If ρ is the representation of D_8 just discussed, then under $\det \rho$ we find that rotations act as the identity and reflections act as multiplication by -1 .

The definition of “representation” is formally very similar to the definition of “action,” and indeed, since every linear automorphism of V is a permutation on the set of vectors in V , the representations of G on V form a subset of the actions of G on V . But the set of representations is in general a much more interesting object. We see here an instance of a general principle: if a set comes equipped with some extra structure (as a vector space comes with the ability to add elements together), then it is a mistake not to make use of that structure; and the more structure the better.

In order to emphasize this point, and to place representations in a very favorable light, let us start by considering the general story of actions of groups on sets. Suppose, then, that G is a group that acts on a set X . For each x , the set of all elements of the form gx , as g ranges over G , is called the *orbit* of x . It is not hard to show that the orbits form a partition of X .

Example. Let G be the dihedral group D_8 acting on the set X of *ordered pairs* of vertices of the square, of which there are sixteen. Then there are three orbits of G on X , namely $\{AA, BB, CC, DD\}$, $\{AB, BA, BC, CB, CD, DC, DA, AD\}$, and $\{AC, CA, BD, DB\}$.

An action of G on X is called *transitive* if there is just one orbit. In other words, it is transitive if for every x and y in X you can find an element g such that $gx = y$. When an action is *not* transitive, we can consider the action of G on each orbit separately, which effectively breaks up the action into a collection of transitive actions on disjoint sets. So in order to study *all* actions of G on sets it suffices to study *transitive* actions; you can think of actions as “molecules” and transitive actions as the “atoms” into which they can be decomposed. We shall see that this idea of *decomposing into objects that cannot be further decomposed* is fundamental to representation theory.

What are the possible transitive actions? A rich source of such actions comes from subgroups H of G . Given a subgroup H of G , a *left coset* of H is a set of the form $\{gh : h \in H\}$, which is commonly denoted by gH . An elementary result in group theory is that the left cosets form a partition of G (as do the right cosets, if you prefer them). There is an obvious action of G on the set of left cosets of H , which we denote by G/H : if g' is an element of G , then it sends the coset gH to the coset $(g'g)H$.

It turns out that every transitive action is of this form! Given a transitive action of G on a set X , choose some $x \in X$ and let H_x be the subgroup of G consisting of all elements h such that $hx = x$. (This set is called the *stabilizer* of x .) Then one can check that the action of G on X is the same¹ as that of G on the left cosets of H_x . For example, the action of D_8 on the first orbit above is isomorphic to the action on the left cosets of the two-element subgroup H generated by a reflection of the square through its diagonal. If we had made a different

1. By “the same” we mean “isomorphic as sets with G -action.” The casual reader may read this as “the same,” while the more careful reader should stop here and work out, or look up, precisely what is meant.

choice of x , for example the point $x' = gx$, then the subgroup of G fixing x' would just be gH_xg^{-1} . This is a so-called *conjugate subgroup*, and it gives a different description of the same orbit, this time as left cosets of gH_xg^{-1} .

It follows that there is a one-to-one correspondence between transitive actions of G and conjugacy classes of subgroups (that is, collections of subgroups conjugate to some given subgroup). If G acts on our original set X in a nontransitive way, then we can break X up into a union of orbits, each of which, as a result of this correspondence, is associated with a conjugacy class of subgroups. This gives us a convenient “bookkeeping” mechanism for describing the action of G on X : just keep track of how many times each conjugacy class of subgroups arises.

Exercise. Check that in the example earlier the three orbits correspond (respectively) to a two-element subgroup R generated by reflection through a diagonal, the trivial subgroup, and another copy of the group R .

This completely solves the problem of how groups act on sets. The internal structure that controls the action is the *subgroup* structure of G .

In a moment we will see the corresponding solution to the problem of how groups act on vector spaces. First, let us just stare at sets for a while and see why, though we have answered our question, we should not feel too happy about it.²

The problem is that the subgroup structure of a group is *just horrible*.

For example, any finite group of order n is a subgroup of the SYMMETRIC GROUP [III.70] S_n (this is “Cayley’s theorem,” which follows by considering the action of G on itself), so in order to list the conjugacy classes of subgroups of the symmetric group S_n one must understand all finite groups of size less than n .³ Or consider the cyclic group $\mathbb{Z}/n\mathbb{Z}$. The subgroups correspond to the divisors of n , a subtle property of n that makes the cyclic groups behave quite differently as n varies. If n is prime, then there are very few subgroups, while if n is a power of 2 there are quite a few. So number theory is involved even if all we want to do is understand the subgroup structure of a group as simple as a cyclic group.

With some relief we now turn our attention back to linear representations. We will see that, just as with actions on sets, one can decompose representations into “atomic” ones. But, by contrast with the case of sets, these atomic representations (called “irreducibles”) turn out to exhibit quite beautiful regularities.

The nice properties of representation theory come largely from the following fact. While elements of the symmetric group S_n can be multiplied together, elements of $\text{GL}(V)$, being matrices, can be *added* as well as multiplied. (But beware: the sum of two elements of $\text{GL}(V)$ is not necessarily an element of $\text{GL}(V)$, because it may not be invertible. It is, however, an element of the endomorphism algebra $\text{End}(V)$. When $V = \mathbb{C}^n$, $\text{End}(V)$ is just the familiar algebra of all $n \times n$ matrices with complex entries, both invertible and not.)

To see the difference it makes to be able to add, consider the cyclic group $G = \mathbb{Z}/n\mathbb{Z}$. For each $\omega \in \mathbb{C}$ with $\omega^n = 1$, we get a representation χ_ω of G on \mathbb{C} by associating the element $r \in \mathbb{Z}/n\mathbb{Z}$ with multiplication by ω^r , which we think of as a linear map from the one-dimensional space \mathbb{C} to itself. This gives us n different one-dimensional representations, one for each n th root of unity, and it turns out that there are no others. Moreover, if $\rho : G \rightarrow \text{GL}(V)$ is any representation of $\mathbb{Z}/n\mathbb{Z}$, then we can write it as a direct sum of these representations by imitating the formula for finding the Fourier mode of a function. Using the representation ρ , we associate with each r in $\mathbb{Z}/n\mathbb{Z}$ a linear map $\rho(r)$. Now let us define a linear map $p_\omega : V \rightarrow V$ by the formula

$$p_\omega = \frac{1}{n} \sum_{0 \leq r < n} \omega^{-r} \rho(r).$$

Then p_ω is an element of $\text{End}(V)$, and one can check that it is actually a PROJECTION [III.52 §3.5] onto a subspace V_ω of V . In fact, this subspace is an EIGENSPACE [I.3 §4.3]: it consists of all vectors v such that $\rho(1)v = \omega v$, which implies, since ρ is a representation, that $\rho(r)v = \omega^r v$. The projection p_ω should be thought of as the analogue of the n th FOURIER COEFFICIENT [III.27] $a_n(f)$ of a function $f(\theta)$ on the circle; note the formal similarity of the above formula to the Fourier expansion formula $a_n(f) = \int e^{-2\pi i n \theta} f(\theta) d\theta$.

Now the interesting thing about the Fourier series of f is that, under favorable circumstances, it adds up to f itself: that is, it decomposes f into TRIGONOMETRIC FUNCTIONS [III.94]. Similarly, what is interesting about the subspaces V_ω is that we can use them to decompose the representation ρ . The composition of any two

PUP: appearance of both 1/16 and 1/6 in the footnote is OK.

2. Exercise: go back to the example of D_8 and list all the possible transitive actions.

3. THE CLASSIFICATION OF FINITE SIMPLE GROUPS [V.8] does at least allow us to estimate the *number* y_n of subgroups of S_n up to conjugacy: it is a result of Pyber that $2^{((1/16)+o(1))n^2} \leq y_n \leq 24^{((1/6)+o(1))n^2}$. Equality is expected for the lower bound.

distinct projections p_ω is 0, from which it can be shown that

$$V = \bigoplus_{\omega} V_{\omega}.$$

We can write each subspace V_{ω} as a sum of one-dimensional spaces, which are copies of \mathbb{C} , and the restriction of ρ to any one of these is just the simple representation χ_{ω} defined earlier. Thus, ρ has been decomposed as a combination of very simple “atoms” χ_{ω} .⁴

This ability to add matrices has a very useful consequence. Let a finite group G act on a complex vector space V . A subspace W of V is called *G-invariant* if $gW = W$ for every $g \in G$. Let W be a G -invariant subspace, and let U be a complementary subspace (that is, one such that every element v of V can be written in exactly one way as $w + u$ with $w \in W$ and $u \in U$). Let ϕ be an arbitrary projection onto U . Then it is a simple exercise to show that the linear map $1/|G| \sum_{g \in G} g\phi$ is also a projection onto a complementary subspace, but with the added advantage that it is G -invariant. This latter fact follows because applying an element g' to the sum just rearranges its terms.

The reason this is so useful is that it allows us to decompose an arbitrary representation into a direct sum of *irreducible representations*, which are representations without a G -invariant subspace. Indeed, if ρ is *not* irreducible, then there is a G -invariant subspace W . By the above remark, we can write $G = W \oplus W'$ with W' also G -invariant. If either W or W' has a further G -invariant subspace, then we can decompose it further, and so on. We have just seen this done for the cyclic group: in that case the irreducible representations were the one-dimensional representations χ_{ω} .

The irreducible representations are the basic building blocks of arbitrary complex representations, just as the basic building blocks for actions on sets are the transitive actions. It raises the question of what the irreducible representations are, a question that has been answered for many important examples, but which is not yet solvable by any general procedure.

To return to the difference between actions and representations, another important observation is that any action of a group G on a finite set X can be *linearized* in the following sense. If X has n elements, then we can

look at the HILBERT SPACE [III.37] $L^2(X)$ of all complex-valued functions defined on X . This has a natural basis given by the “delta functions” δ_x , which send x to 1 and all other elements of X to 0. Now we can turn the action of G on X into an action of G on the basis in an obvious way: we just define $g\delta_x$ to be δ_{gx} . We can extend this definition by linearity, since an arbitrary function f is a linear combination of the basis functions δ_x . This gives us an action of G on $L^2(X)$, which can be defined by a simple formula: if f is a function in $L^2(X)$, then gf is the function defined by $(gf)(x) = f(g^{-1}x)$. Equivalently, gf does to gx what f does to x . Thus, an action on sets can be thought of as an assignment of a very special matrix to every group element, namely a matrix with only 0s and 1s and precisely one 1 in each row and each column. (Such matrices are called *permutation matrices*.) By contrast, a general representation assigns an *arbitrary* invertible matrix.

Now, even when X itself is a single orbit under the action of G , the above representation on $L^2(X)$ can break up into pieces. For an extreme example of this phenomenon, consider the action of $\mathbb{Z}/n\mathbb{Z}$ on itself by multiplication. We have just seen that, by means of the “Fourier expansion” above, this breaks up into a sum of n one-dimensional representations.

Let us now consider the action of an arbitrary group G on itself by multiplication, or, to be more precise, left multiplication. That is, we shall associate with each element g the permutation of G that takes each h in G to gh . This action is obviously transitive. As an action on a *set* it cannot be decomposed any further. But when we *linearize* this action to a representation of G on the vector space $L^2(G)$, we have much greater flexibility to decompose the action. It turns out that, not only does it break up into a direct sum of many irreducible representations, but *every* irreducible representation ρ of G occurs as one of the summands in this direct sum, and the number of times that ρ appears is equal to the dimension of the subspace on which it acts.

The representation we have just discussed is called the *left regular representation* of G . The fact that every irreducible representation occurs in it so regularly makes it extremely useful. Notice that it is easier to decompose representations on complex vector spaces than on real vector spaces, since every automorphism of a complex vector space has an eigenvector. So it is simplest to begin by studying complex representations.

The time has now come to state the fundamental theorem about complex representations of finite groups.

4. To summarize the rest of this article: the similarity to the Fourier transform is not just analogy—decomposing a representation into its irreducible summands is a notion that includes both this example and the Fourier transform.

This theorem tells us how many irreducible representations there are for a finite group, and, more colorfully, that representation theory is a “non-Abelian analogue of Fourier decomposition.”

Let $\rho : G \rightarrow \text{End}(V)$ be a representation of G . The *character* χ_ρ of ρ is defined to be its trace: that is, χ_ρ is a function from G to \mathbb{C} and $\chi_\rho(g) = \text{tr}(\rho(g))$ for each g in G . Since $\text{tr}(AB) = \text{tr}(BA)$ for any two matrices A and B , we have $\chi_\rho(hgh^{-1}) = \chi_\rho(g)$. Therefore, χ_ρ is very far from an arbitrary function on G : it is a function that is constant on each *conjugacy class*. Let K_G denote the vector space of all complex-valued functions on G with this property; it is called the *representation ring* of G .

The characters of the irreducible representations of a group form a very important set of data about the group, which it is natural to organize into a matrix. The columns are indexed by the conjugacy classes, the rows by the irreducible representations, and each entry is the value of the character of the given representation at the given conjugacy class. This array is called the *character table* of the group, and it contains all the important information about representations of the group: it is our periodic table. The basic theorem of the subject is that this array is a *square*.

Theorem (the character table is square). *Let G be a finite group. Then the characters of the irreducible representations form an orthonormal basis of K_G .*

When we say that the basis of characters is *orthonormal* we mean that the Hermitian inner product defined by

$$\langle \chi, \psi \rangle = |G|^{-1} \sum_{g \in G} \chi(g) \overline{\psi(g)}$$

is 1 when $\chi = \psi$ and 0 otherwise. The fact that it is a basis implies in particular that there are exactly as many irreducible representations as there are conjugacy classes in G , and the map from isomorphism classes of representations to K_G that sends each ρ to its character is an injection. That is, an arbitrary representation is determined up to isomorphism by its character.

The internal structure of a group G that controls how it can act on vector spaces is the structure of conjugacy classes of elements of G . This is a much gentler structure than the set of all conjugacy classes of *subgroups* of G . For example, in the symmetric group S_n two permutations belong to the same conjugacy class if and only if they have the same cycle type. Therefore,

in that group there is a bijection between conjugacy classes and partitions of n .⁵

Furthermore, whereas it is completely unclear how to count subgroups, conjugacy classes are much easier to handle. For instance, since they partition the group, we have the formula $|G| = \sum_{C \text{ a conjugacy class}} |C|$. On the representation side, there is a similar formula, which arises from the decomposition of the regular representation $L^2(G)$ into irreducibles: $|G| = \sum_{V \text{ irreducible}} (\dim V)^2$. It is inconceivable that there might be a similarly simple formula for sums over all subgroups of a group.

We have reduced the problem of understanding the general structure of the representations of a finite group G to the problem of determining the character table of G . When $G = \mathbb{Z}/n\mathbb{Z}$, our description of the n irreducible representations above implies that all the entries of this matrix are roots of unity. Here are the character tables for D_8 (on the left), the group of symmetries of the square, and, just for contrast, for the group $\mathbb{Z}/3\mathbb{Z}$ (on the right):

1	1	1	1	1	1	1	1
1	1	1	-1	-1	1	z	z^2
1	1	-1	1	-1	1	z^2	z
1	1	-1	-1	1			
2	-2	0	0	0			

where $z = \exp(2\pi i/3)$.

The obvious question—Where did the first table come from?—indicates the main problem with the theorem: though it tells us the shape of the character table, it leaves us no closer to understanding what the actual character values are. We know *how many* representations there are, but not *what* they are, or even what their dimensions are. We do not have a general method for constructing them, a kind of “non-Abelian Fourier transform.” This is the central problem of representation theory.

Let us see how this problem can be solved for the group D_8 . Over the course of this article, we have already encountered three irreducible representations of this group. The first is the “trivial” one-dimensional representation: the homomorphism $\rho : D_8 \rightarrow \text{GL}_1$ that takes every element of D_8 to the identity. The second is the two-dimensional representation we wrote down in the first section, where each element of D_8 acts on \mathbb{R}^2

5. Not only is the set of all partitions a sensible combinatorial object, it is far smaller than the set of all subgroups of S_n : HARDY [VI.73] and RAMANUJAN [VI.82] showed that the number of partitions of n is about $(1/4n\sqrt{3})e^{\pi\sqrt{(2n/3)}}$.

in the obvious way. The determinant of this representation is a one-dimensional representation that is *not* trivial: it sends the rotations to 1 and the reflections to -1 . So we have constructed the first three rows of the character table above. There are five conjugacy classes in D_8 (trivial, reflection through axis, reflection through diagonal, 90° rotation, 180° rotation), so we know that there are just two more rows.

The equality $|G| = 8 = 2^2 + 1 + 1 + (\dim V_4)^2 + (\dim V_5)^2$ implies that these missing representations are one dimensional. One way of getting the missing character values is to use orthogonality of characters.

A slightly (but only slightly) less ad hoc way is to decompose $L^2(X)$ for small X . For example when X is the pair of diagonals $\{AC, BD\}$, we have $L^2(X) = V_4 \oplus \mathbb{C}$, where \mathbb{C} is the trivial representation.

We are now going to start pointing the way toward some more modern topics in representation theory. Of necessity, we will use language from fairly advanced mathematics: the reader who is familiar with only some of this language should consider browsing the remaining sections, since different discussions have different prerequisites.

In general, a good, but not systematic, way of finding representations is to find objects on which G acts, and “linearize” the action. We have seen one example of this: when G acts on a set X we can consider the linearized action on $L^2(X)$. Recall that the irreducible G -sets are all of the form G/H , for H some subgroup of G . As well as looking at $L^2(G/H)$, we can consider, for every representation W of H , the vector space $L^2(G/H, W) = \{f : G \rightarrow W \mid f(gh) = h^{-1}f(g), g \in G, h \in H\}$; in geometric language, for those who prefer it, this is the space of sections of the associated W -bundle on G/H . This representation of G is called the *induced representation* of W from H to G .

Other linearizations are also important. For example, if G acts continuously on a topological space X , we can consider how it acts on homology classes and hence on the HOMOLOGY GROUPS [IV.6 §4] of X .⁶ The simplest case of this is the map $z \rightarrow \bar{z}$ of the circle S^1 . Since this map squares to the identity map, it gives us an action of $\mathbb{Z}/2\mathbb{Z}$ on S^1 , which becomes a representation of $\mathbb{Z}/2\mathbb{Z}$ on $H_1(S^1) = \mathbb{R}$ (which represents the identity as multiplication by 1 and the other element of $\mathbb{Z}/2\mathbb{Z}$ as multiplication by -1).

6. The homology groups discussed in the article just referred to consist of formal sums of homology classes with integer coefficients. Here, where a vector space is required, we are taking real coefficients.

Methods like these have been used to determine the character tables of all finite SIMPLE GROUPS [I.3 §3.3], but they still fall short of a uniform description valid for all groups.

There are many arithmetic properties of the character table that hint at properties of the desired non-Abelian Fourier transform. For example, the size of a conjugacy class divides the order of the group, and in fact the dimension of a representation also divides the order of the group. Pursuing this thought leads to an examination of the values of the characters mod p , relating them to the so-called *p-local subgroups*. These are groups of the form $N(Q)/Q$, where Q is a subgroup of G , the number of elements of Q is a power of p , and $N(Q)$ is the *normalizer* of Q (defined to be the largest subgroup of G that contains Q as a normal subgroup). When the so-called “*p*-Sylow subgroup” of G is Abelian, beautiful conjectures of Broué give us an essentially complete picture of the representations of G . But in general these questions are at the center of a great deal of contemporary research.

3 Fourier Analysis

We have justified the study of group actions on vector spaces by explaining that the theory of representations has a nice structure that is not present in the theory of group actions on sets. A more historically based account would start by saying that spaces of functions very often come with natural actions of some group G , and many problems of traditional interest can be related to the decomposition of these representations of G .

In this section we will concentrate on the case where G is a compact LIE GROUP [III.50 §1]. We will see that in this case many of the nice features of the representation theory of finite groups persist.

The prototypical example is the space $L^2(S^1)$ of square-integrable functions on the circle S^1 . We can think of the circle as the unit circle in \mathbb{C} , and thereby identify it with the group of rotations of the circle (since multiplication by $e^{i\theta}$ rotates the circle by θ). This action linearizes to an action on $L^2(S^1)$: if f is a square-integrable function defined on S^1 and w belongs to the circle, then $(w \cdot f)(z)$ is defined to be $f(w^{-1}z)$. That is, $w \cdot f$ does to wz what f does to z .

Classical Fourier analysis expands functions in $L^2(S^1)$ in terms of a basis of trigonometric functions: the functions z^n for $n \in \mathbb{Z}$. (These look more “trigonometric” if one writes $e^{i\theta}$ for z and $e^{in\theta}$ for z^n .) If we

fix w and write $\phi_n(z) = z^n$, then $(w \cdot \phi_n)(z) = \phi_n(w^{-1}z) = w^{-n}\phi_n(z)$. In particular, $w \cdot \phi_n$ is a multiple of ϕ_n for each w , so the one-dimensional subspace generated by ϕ_n is invariant under the action of S^1 . In fact, every irreducible representation of S^1 is of this form, as long as we restrict attention to continuous representations.

Now let us consider an innocuous-looking generalization of the above situation: we shall replace 1 by n and try to understand $L^2(S^n)$, the space of complex-valued square-integrable functions on the n -sphere S_n . The n -sphere is acted on by the group of rotations $\text{SO}(n+1)$. As usual, this can be converted into a representation of $\text{SO}(n+1)$ on the space $L^2(S^n)$, which we would like to decompose into irreducible representations; equivalently, we would like to decompose $L^2(S^n)$ into a direct sum of minimal $\text{SO}(n+1)$ -invariant subspaces.

This turns out to be possible, and the proof is very similar to the proof for finite groups. In particular, a compact group such as $\text{SO}(n+1)$ has a natural PROBABILITY MEASURE [III.73 §2] on it (called *Haar measure*) in terms of which we can define averages. Roughly speaking, the only difference between the proof for $\text{SO}(n+1)$ and the proof in the finite case is that we have to replace a few sums by integrals.

The general result that one can prove by this method is the following. If G is a compact group that acts continuously on a compact space X (in the sense that each permutation $\phi(g)$ of X is continuous, and also that $\phi(g)$ varies continuously with g), then $L^2(X)$ splits up into an orthogonal direct sum of finite-dimensional minimal G -invariant subspaces; equivalently, the linearized action of G on $L^2(X)$ splits up into an orthogonal direct sum of irreducible representations, all of which are finite dimensional. The problem of finding a Hilbert space basis of $L^2(X)$ then splits into two subproblems: we must first determine the irreducible representations of G , a problem which is independent of X , and then determine how many times each of these irreducible representations occurs in $L^2(X)$.

When $G = S^1$ (which we identified with $\text{SO}(2)$) and $X = S^1$ as well, we saw that these irreducible representations were one dimensional. Now let us look at the action of the compact group $\text{SO}(3)$ on S^2 . It can be shown that the action of G on $L^2(S^2)$ commutes with the *Laplacian*, the differential operator Δ on $L^2(S^2)$ defined by

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}.$$

That is, $g(\Delta f) = \Delta(gf)$ for any $g \in G$ and any (sufficiently smooth) function f . In particular, if f is an eigenfunction for the Laplacian (which means that $\Delta f = \lambda f$ for some $\lambda \in \mathbb{C}$), then for each $g \in \text{SO}(3)$ we have

$$\Delta gf = g\Delta f = g\lambda f = \lambda gf,$$

so gf is also an eigenfunction for Δ . Therefore, the space V_λ of all eigenvectors for the Laplacian with eigenvalue λ is G -invariant. In fact, it turns out that if V_λ is nonzero then the action of G on V_λ is an irreducible representation. Furthermore, each irreducible representation of $\text{SO}(3)$ arises exactly once in this way. More precisely, we have a Hilbert space direct sum,

$$L^2(S^2) = \bigoplus_{n \geq 0} V_{2n(2n+2)},$$

and each eigenspace $V_{2n(2n+2)}$ has dimension $2n+1$. Note that this is a case where the set of eigenvalues is *discrete*. (These eigenspaces are discussed further in SPHERICAL HARMONICS [III.89].)

The nice feature that each irreducible representation appears at most once is rather special to the example $L^2(S^n)$. (For an example where this does not happen, recall that with the regular representation $L^2(G)$ of a finite group G each irreducible representation ρ occurs $\dim \rho$ times in $L^2(G)$.) However, other features are more generic: for example, when a compact Lie group acts differentiably on a space X , then the sum of all the G -invariant subspaces of $L^2(X)$ corresponding to a particular representation is always equal to the set of common eigenvectors of some family of commuting differential operators. (In the example above, there was just one operator, the Laplacian.)

Interesting SPECIAL FUNCTIONS [III.87], such as solutions of certain differential equations, often admit representation-theoretic meaning, for example as matrix coefficients. Their properties can then easily be deduced from general results in functional analysis and representation theory rather than from any calculation. Hypergeometric equations, Bessel equations, and many integrable systems arise in this way.

There is more to say about the similarities between the representation theory of compact groups and that of finite groups. Given a compact group G and an irreducible representation ρ of G , we can again take its trace (since it is finite dimensional) and thereby define its character χ_ρ . Just as before, χ_ρ is constant on each conjugacy class. Finally, “the character table is square,” in the sense that the characters of the irreducible representations form an orthonormal basis of

the Hilbert space of all square-integrable functions that are conjugation invariant in this sense. (Now, though, the “square matrix” is infinite.) When $G = S^1$ this is the Fourier theorem; when G is finite this is the theorem of section 2.

4 Noncompact Groups, Groups in Characteristic p , and Lie Algebras

The “character table is square” theorem focuses our attention on groups with nice conjugacy-class structure. What happens when we take such a group but relax the requirement that it be compact?

A paradigmatic noncompact group is the real numbers \mathbb{R} . Like S^1 , \mathbb{R} acts on itself in an obvious way (the real number t is associated with the translation $s \mapsto s + t$), so let us linearize that action in the usual way and look for a decomposition of $L^2(\mathbb{R})$ into \mathbb{R} -invariant subspaces.

In this situation we have a *continuous family* of irreducible one-dimensional representations: for each real number λ we can define the function χ_λ by $\chi_\lambda(x) = e^{2\pi i \lambda x}$. These functions are not square integrable, but despite this difficulty classical Fourier analysis tells us that we can write an L^2 -function in terms of them. However, since the Fourier modes now vary in a continuous family, we can no longer decompose a function as a sum: rather we must use an integral. First, we define the Fourier transform \hat{f} of f by the formula $\hat{f}(\lambda) = \int f(x) e^{2\pi i \lambda x} dx$. The desired decomposition of f is then $f(x) = \int \hat{f}(\lambda) e^{-2\pi i \lambda x} d\lambda$. This, the *Fourier inversion formula*, tells us that f is a weighted integral of the functions χ_λ . We can also think of it as something like a decomposition of $L^2(\mathbb{R})$ as a “direct integral” (rather than direct sum) of the one-dimensional subspaces generated by the functions χ_λ . However, we must treat this picture with due caution since the functions χ_λ do not belong to $L^2(\mathbb{R})$.

This example indicates what we should expect in general. If X is a space with a measure and G acts continuously on it in a way that preserves the measures of subsets of X (as translations did with subsets of \mathbb{R}), then the action of G on X gives rise to a measure μ_X defined on the set of all irreducible representations, and $L^2(X)$ can be decomposed as the integral over all irreducible representations with respect to this measure. A theorem that explicitly describes such a decomposition is called a *Plancherel* theorem for X .

For a more complicated but more typical example, let us look at the action of $SL_2(\mathbb{R})$ (the group of real

2×2 matrices with determinant 1) on \mathbb{R}^2 and see how to decompose $L^2(\mathbb{R}^2)$. As we did when we looked at functions defined on S^2 , we shall make use of a differential operator. This involves the small technicality that we should look at smooth functions, and we do not ask for them to be defined at the origin. The appropriate differential operator this time turns out to be the Euler vector field $x(\partial/\partial x) + y(\partial/\partial y)$. It is not hard to check that if f satisfies the condition $f(tx, ty) = t^s f(x, y)$ for every x, y , and $t > 0$, then f is an eigenfunction of this operator with eigenvalue s , and indeed all functions in the eigenspace with this eigenvalue, which we shall denote by W_s , are of this form. We can also split W_s up as $W_s^+ \oplus W_s^-$, where W_s^+ and W_s^- consist of the even and odd functions in W_s , respectively.

The easiest way of analyzing the structure of W_s is to compute the action of the LIE ALGEBRA [III.50 §2] \mathfrak{sl}_2 . For those readers unfamiliar with Lie algebras, we will say only that the Lie algebra of a Lie group G keeps track of the action of elements of G that are “infinitesimally close to the identity,” and that in this case the Lie algebra \mathfrak{sl}_2 can be identified with the space of 2×2 matrices of trace 0, with $\begin{pmatrix} a & b \\ c & -a \end{pmatrix}$ acting as the differential operator $(-ax - by)(\partial/\partial x) + (-cx + ay)(\partial/\partial y)$.

Every element of W_s is a function on \mathbb{R}^2 . If we restrict these functions to the unit circle, then we obtain a map from W_s to the space of smooth functions defined on S^1 , which turns out to be an isomorphism. We already know that this space has a basis of Fourier modes z^m , which we can now think of as $(x + iy)^m$, defined when $x^2 + y^2 = 1$. There is a unique extension of this from a function defined on S^1 to a function in W_s , namely the function $w_m(x, y) = (x + iy)^m (x^2 + y^2)^{(s-m)/2}$. One can then check the following actions of simple matrices on these functions (to do so, recall the association of the matrices with differential operators given in the previous paragraph):

$$\begin{aligned} \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \cdot w_m &= m w_m, \\ \begin{pmatrix} 1 & i \\ i & -1 \end{pmatrix} \cdot w_m &= (m - s) w_{m+2}, \\ \begin{pmatrix} 1 & -i \\ -i & -1 \end{pmatrix} \cdot w_m &= (-m - s) w_{m-2}. \end{aligned}$$

It follows that if s is not an integer, then from any function w_m in W_s^+ we can produce all the others using the action of $SL_2(\mathbb{R})$. Therefore, $SL_2(\mathbb{R})$ acts irreducibly on W_s^+ . Similarly, it acts irreducibly on W_s^- . We have therefore encountered a significant difference between

this and the finite/compact case: when G is not compact, irreducible representations of G can be infinite dimensional.

Looking more closely at the formulas for W_s when $s \in \mathbb{Z}$, we see more disturbing differences. In order to understand these, let us distinguish carefully between representations that are *reducible* and representations that are *decomposable*. The former are representations that have nontrivial G -invariant subspaces, whereas the latter are representations where one can decompose the space on which G acts into a direct sum of G -invariant subspaces. Decomposable representations are obviously reducible. In the finite/compact case, we used an averaging process to show that reducible representations are decomposable. Now we do not have a natural probability measure to use for the averaging, and it turns out that there can be reducible representations that are not decomposable.

Indeed, if s is a nonnegative integer, then the subspaces W_s^+ and W_s^- give us an example of this phenomenon. They are indecomposable (in fact, this is true even when s is a negative integer not equal to -1) but they contain an invariant subspace of dimension $s + 1$. Thus, we cannot write the representation as a direct sum of irreducible representations. (One can do something a little bit weaker, however: if we quotient out by the $(s + 1)$ -dimensional subspace, then the quotient representation can be decomposed.)

It is important to understand that in order to produce these indecomposable but reducible representations we worked not in the space $L^2(\mathbb{R}^2)$ but in the space of smooth functions on \mathbb{R}^2 with the origin removed. For instance, the functions w_m above are not square integrable. If we look just at representations of G that act on subspaces of $L^2(X)$, then we *can* split them up into a direct sum of irreducibles: given a G -invariant subspace, its orthogonal complement is also G -invariant. It might therefore seem best to ignore the other, rather subtle representations and just look at these ones. But it turns out to be easier to study *all* representations and only later ask which ones occur inside $L^2(X)$. For $\mathrm{SL}_2(\mathbb{R})$, the representations we have just constructed (which were subquotients of W_s^\pm) exhaust all the irreducible representations,⁷ and there is a Plancherel formula for $L^2(\mathbb{R}^2)$ that tells us which ones appear in

$L^2(\mathbb{R}^2)$ and with what multiplicity:

$$L^2(\mathbb{R}^2) = \int_{-\infty}^{\infty} W_{-1+it} e^{it} dt.$$

To summarize: if G is not compact, then we can no longer take averages over G . This has various consequences:

Representations occur in continuous families. The decomposition of $L^2(X)$ takes the form of a direct integral, not a direct sum.

Representations do not split up into a direct sum of irreducibles. Even when a representation admits a finite composition series, as with the action of $\mathrm{SL}^2(\mathbb{R})$ on W_s^\pm , it need not split up into a direct sum. So to describe all representations we need to do more than just describe the irreducibles—we also need to describe the glue that holds them together.

So far, the theory of representations of a noncompact group G seems to have *none* of the pleasant features of the compact case. But one thing does survive: there is still an analogue of the theorem that the character table is square. Indeed, we can still define characters in terms of the traces of group elements. But now we must be careful, since the irreducible representation may be on an infinite-dimensional vector space, so that its trace cannot be defined so easily. In fact, characters are not functions on G , but only DISTRIBUTIONS [III.18]. The character of a representation determines the *semisimplification* of a representation ρ : that is, it tells us which irreducible representations are part of ρ , but not how they are glued together.⁸

These phenomena were discovered by Harish-Chandra in the 1950s in an extraordinary series of works that completely described the representation theory of Lie groups such as the ones we have discussed (the precise condition is that they should be real and reductive—a concept that will be explained later in this article) and the generalizations of classical theorems of Fourier analysis to this setting.⁹

Independently and slightly earlier, Brauer had investigated the representation theory of *finite* groups on finite-dimensional vector spaces over fields of characteristic p . Here, too, reducible representations need not decompose as direct sums, though in this case the

7. To make this precise requires some care about what we mean by “isomorphic.” Because many different topological vector spaces can have the same underlying \mathfrak{sl}_2 -module, the correct notion is of *infinitesimal* equivalence. Pursuing this notion leads to the category of *Harish-Chandra modules*, a category with good finiteness properties.

8. It is a major theorem of Harish-Chandra that the distribution that defines a character is given by *analytic* functions on a dense subset of the semisimple elements of the group.

9. The problem of determining the irreducible *unitary* representations for real reductive groups has still not been solved; the most complete results are due to Vogan.

problem is not lack of compactness (obviously, since everything is finite) but an inability to *average* over the group: we would like to divide by $|G|$, but often this is zero. A simple example that illustrates this is the action of $\mathbb{Z}/p\mathbb{Z}$ on the space \mathbb{F}_p^2 that takes x to the 2×2 matrix $\begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix}$. This is reducible, since the column vector $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ is fixed by the action, and therefore generates an invariant subspace. However, if one could decompose the action, then the matrices $\begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix}$ would all be diagonalizable, which they are not.

It is possible for there to be infinitely many indecomposable representations, which again may vary in families. However, as before, there are only finitely many *irreducible* representations, so there is some chance of a “character table is square” theorem in which the rows of the square are parametrized by characters of irreducible representations. Brauer proved just such a theorem, pairing the characters with *p-semisimple* conjugacy classes in G : that is, conjugacy classes of elements whose order is not divisible by p .

We will draw two crude morals from the work of Harish-Chandra and of Brauer. The first is that the category of representations of a group is always a reasonable object, but when the representations are infinite dimensional it requires serious technical work to set it up. Objects in this category do not necessarily decompose as a direct sum of irreducibles (one says that the category is not *semisimple*), and can occur in infinite families, but irreducible objects pair off in some precise way with certain “diagonalizable” conjugacy classes in the group—there is always some kind of analogue of “the character table is square” theorem.

It turns out that when we consider representations in more general contexts—Lie algebras acting on vector spaces, quantum groups, p -adic groups on infinite-dimensional complex or p -adic vector spaces, etc.—these qualitative features stay the same.

The second moral is that we should always hope for some “non-Abelian Fourier transform”: that is, a set that parametrizes irreducible representations and a description of the character values in terms of this set.

In the case of real reductive groups Harish-Chandra’s work provides such an answer, generalizing the Weyl character formula for compact groups; for arbitrary groups no such answer is known. For special classes of groups, there are partially successful general principles (the orbit method, Broué’s conjecture), of which the deepest are the extraordinary circle of conjec-

tures known as the Langlands program, which we shall discuss later.

5 Interlude: The Philosophical Lessons of “The Character Table Is Square”

Our basic theorem (“the character table is square”) tells us to expect that the category of all irreducible representations of G is interesting when the conjugacy-class structure of G is in some way under control. We will finish this essay by explaining a remarkable family of examples of such groups—the rational points of *reductive* algebraic groups—and their conjectured representation theory, which is described by the *Langlands program*.

An *affine algebraic group* is a subgroup of some group GL_n that is defined by polynomial equations in the matrix coefficients. For example, the determinant of a matrix is a polynomial in the matrix coefficients, so the group SL_n , which consists of all matrices in GL_n with determinant 1, is such a group. Another is SO_n , which is the set of matrices with determinant 1 that satisfy the equation $AA^T = I$.

The above notation did not specify what sort of coefficients we were allowing for the matrices. That vagueness was deliberate. Given an algebraic group G and a field k , let us write $G(k)$ for the group where the coefficients are taken to have values in k . For example, $\mathrm{SL}_n(\mathbb{F}_q)$ is the set of $n \times n$ matrices with coefficients in the finite field \mathbb{F}_q and determinant 1. This group is finite, as is $\mathrm{SO}_n(\mathbb{F}_q)$, while $\mathrm{SL}_n(\mathbb{R})$ and $\mathrm{SO}_n(\mathbb{R})$ are Lie groups. Moreover, $\mathrm{SO}_n(\mathbb{R})$ is compact, while $\mathrm{SL}_n(\mathbb{R})$ is not. So among affine algebraic groups over fields one already finds all three types of groups we have discussed: finite groups, compact Lie groups, and noncompact Lie groups.

We can think of $\mathrm{SL}_n(\mathbb{R})$ as the set of matrices in $\mathrm{SL}_n(\mathbb{C})$ that are equal to their complex conjugates. There is another involution on $\mathrm{SL}_n(\mathbb{C})$ that is a sort of “twisted” form of complex conjugation, where we send a matrix A to the complex conjugate of $(A^{-1})^T$. The fixed points of this new involution (that is, the determinant-1 matrices A such that A equals the complex conjugate of $(A^{-1})^T$) form a group called $\mathrm{SU}_n(\mathbb{R})$. This is also called a *real form* of $\mathrm{SL}_n(\mathbb{C})$,¹⁰ and it is compact.

10. When we say that $\mathrm{SL}_n(\mathbb{R})$ and $\mathrm{SU}_n(\mathbb{R})$ are both “real forms” of $\mathrm{SL}_n(\mathbb{C})$, what is meant more precisely is that in both cases the group can be described as a subgroup of some group of real matrices that consists of all solutions to a set of polynomial equations, and that when the same set of equations is applied instead to the group of complex matrices the result is isomorphic to $\mathrm{SL}_n(\mathbb{C})$.

The groups $\mathrm{SL}_n(\mathbb{F}_q)$ and $\mathrm{SO}_n(\mathbb{F}_q)$ are almost simple groups;¹¹ the classification of finite simple groups tells us, mysteriously, that all but twenty-six of the finite simple groups are of this form. A much, much easier theorem tells us that the *connected compact* groups are also of this form.

Now, given an algebraic group G , we can also consider the instances $G(\mathbb{Q}_p)$, where \mathbb{Q}_p is the field of p -adic numbers, and also $G(\mathbb{Q})$. For that matter, we may consider $G(k)$ for any other field k , such as the FUNCTION FIELD OF AN ALGEBRAIC VARIETY [V.33]. The lesson of section 4 is that we may hope for all of these many groups to have a good representation theory, but that to obtain it there will be serious “analytic” or “arithmetic” difficulties to overcome, which will depend strongly on the properties of the field k .

Lest the reader adopt too optimistic a viewpoint, we point out that not every affine algebraic group has a nice conjugacy-class structure. For example, let V_n be the set of upper triangular matrices in GL_n with 1s along the diagonal, and let k be \mathbb{F}_q . For large n , the conjugacy classes in $V_n(\mathbb{F}_q)$ form large and complex families: to parametrize them sensibly one needs more than n parameters (in other words, they belong to families of dimension greater than n , in an appropriate sense), and it is not in fact known how to parametrize them even for a smallish value of n , such as 11. (It is not obvious that this is a “good” question though.)

More generally, solvable groups tend to have horrible conjugacy-class structure, even when the groups themselves are “sensible.” So we might expect their representation theory to be similarly horrible. The best we can hope for is a result that describes the entries of the character table *in terms of* this horrible structure—some kind of non-Abelian Fourier integral. For certain p -groups Kirillov found such a result in the 1960s, as an example of the “orbit method,” but the general result is not yet known.

On the other hand, groups that are similar to connected compact groups do have a nice conjugacy-class structure: in particular, finite simple groups do. An algebraic group is called *reductive* if $G(\mathbb{C})$ has a compact real form. So, for instance, SL_n is reductive by the existence of the real form $\mathrm{SU}_n(\mathbb{R})$. The groups GL_n and SO_n are also reductive, but V_n is not.¹²

Let us examine the conjugacy classes in the group SU_n . Every matrix in $\mathrm{SU}_n(\mathbb{R})$ can be diagonalized, and two conjugate matrices have the same eigenvalues, up to reordering. Conversely, any two matrices in $\mathrm{SU}_n(\mathbb{R})$ with the same eigenvalues are conjugate. Therefore, the conjugacy classes are parametrized by the quotient of the subgroup of all diagonal matrices by the action of S_n that permutes the entries.

This example can be generalized. Any compact connected group has a *maximal torus* T , that is, a maximal subgroup isomorphic to a product of circles. (In the previous example it was the subgroup of diagonal matrices.) Any two maximal tori are conjugate in G , and any conjugacy class in G intersects T in a unique W -orbit on T , where W is the *Weyl group*, the finite group $N(T)/T$ (where $N(T)$ is the normalizer of T).

The description of conjugacy classes in $G(\bar{k})$, for an algebraically closed field \bar{k} , is only a little more complicated. Any element $g \in G(\bar{k})$ admits a JORDAN DECOMPOSITION [III.45]: it can be written as $g = su = us$, where s is conjugate to an element of $T(\bar{k})$ and u is unipotent when considered as an element of $\mathrm{GL}_n(\bar{k})$. (A matrix A is *unipotent* if some power of $A - I$ is zero.) Unipotent elements never intersect compact subgroups. When $G = \mathrm{GL}_n$ this is the usual Jordan decomposition; conjugacy classes of unipotent elements are parametrized by partitions of n , which, as we mentioned in section 2, are precisely the conjugacy classes of $W = S_n$. For general reductive groups, unipotent conjugacy classes are again almost the same thing as conjugacy classes in W .¹³ In particular, there are finitely many, independent of \bar{k} .

Finally, when k is not algebraically closed, one describes conjugacy classes by a kind of Galois descent; for example, in $\mathrm{GL}_n(k)$, semisimple classes are still determined by their characteristic polynomial, but the fact that this polynomial has coefficients in k constrains the possible conjugacy classes.

The point of describing the conjugacy-class structure in such detail is to describe the representation theory in analogous terms. A crude feature of the conjugacy-class structure is the way it decouples the field k from finite combinatorial data that is attached to G but independent of k —things like W , the lattice defining T , roots, and weights.

11. Which is to say that the quotient of these groups by their center is simple.

12. The miracle, not relevant for this discussion, is that compact connected groups can be easily classified. Each one is essentially a product of circles and non-Abelian simple compact groups. The latter

are parametrized by DYNKIN DIAGRAMS [III.50 §3]. They are SU_n , Sp_{2n} , SO_n , and five others, denoted E_6 , E_7 , E_8 , F_4 , and G_2 . That is it!

13. They are different, but related. Precisely, they are given by combinatorial data, Lusztig’s *two-sided cells* for the corresponding affine Weyl group.

The “philosophy” suggested by the theorem that the character table is square suggests that the representation theory should also admit such a decoupling: it should be built out of the representation theory of k^* , which is the analogue of the circle, and out of the combinatorial structure of $G(\bar{k})$ (such as the finite groups W). Moreover, representations should have a “Jordan decomposition”:¹⁴ the “unipotent” representations should have some kind of combinatorial complexity but little dependence on k , and compact groups should have no unipotent representations.

The Langlands program provides a description along the lines laid out above, but it goes beyond any of the results we have suggested in that it also describes the entries of the character table. Thus, for this class of examples, it gives us (conjecturally) the hoped-for “non-Abelian Fourier transform.”

6 Coda: The Langlands Program

And so we conclude by just hinting at statements. If $G(k)$ is a reductive group, we want to describe an appropriate category of representations for $G(k)$, or at least the character table, which we may think of as a “semisimplification” of that category.

Even when k is finite, it is too much to hope that conjugacy classes in $G(k)$ parametrize irreducible representations. But something not so far off is conjectured, as follows.

To a reductive group G over an algebraically closed field, Langlands attaches another reductive group ${}^L G$, the *Langlands dual*, and conjectures that representations of $G(k)$ will be parametrized by conjugacy classes in ${}^L G(\mathbb{C})$.¹⁵ However, these are not conjugacy classes of *elements* of ${}^L G(\mathbb{C})$, as before, but of *homomorphisms* from the Galois group of k to ${}^L G$. The Langlands dual was originally defined in a combinatorial manner, but there is now a conceptual definition. A few examples of pairs $(G, {}^L G)$ are $(\mathrm{GL}_n, \mathrm{GL}_n)$, $(\mathrm{SO}_{2n+1}, \mathrm{Sp}_{2n})$, and $(\mathrm{SL}_n, \mathrm{PGL}_n)$.

In this way the Langlands program describes the representation theory as built out of the structure of G and the arithmetic of k .

Though this description indicates the flavor of the conjectures, it is not quite correct as stated. For instance, one has to modify the Galois group¹⁶ in such a way that the correspondence is true for the group $\mathrm{GL}_1(k) = k^*$. When $k = \mathbb{R}$, we get the representation theory of \mathbb{R}^* (or its compact form S^1), which is Fourier analysis; on the other hand, when k is a p -adic local field, the representation theory of k^* is described by local class field theory. We already see an extraordinary aspect of the Langlands program: it precisely unifies and generalizes harmonic analysis and number theory.

The most compelling conjectural versions of the Langlands program are “equivalences of derived categories” between the category of representations and certain geometric objects on the spaces of Langlands parameters. These conjectural statements are the hoped-for Fourier transforms.

Though much progress has been made, a large part of the Langlands program remains to be proved. For finite reductive groups, slightly weaker statements have been proved, mostly by Lusztig. As all but twenty-six of the finite simple groups arise from reductive groups, and as the sporadic groups have had their character tables computed individually, this work already determines the character tables of all the finite simple groups.

For groups over \mathbb{R} , the work of Harish-Chandra and later authors again confirms the conjectures. But for other fields, only fragmentary theorems have been proved. There is much still to be done.

Further Reading

A nice introductory text on representation theory is Alperin’s *Local Representation Theory* (Cambridge University Press, Cambridge, 1993). As for the Langlands program, the 1979 American Mathematical Society volume titled *Automorphic Forms, Representations, and L-functions* (but universally known as “The Corvallis Proceedings”) is more advanced, and as good a place to start as any.

IV.10 Geometric and Combinatorial Group Theory

Martin R. Bridson

14. The first such theorems were proved for $\mathrm{GL}_n(\mathbb{F}_q)$ by Green and Steinberg. However, the notion of Jordan decomposition for characters originates with Brauer, in his work on modular representation theory. It is part of his modular analogue of the “character table is square” theorem, which we mentioned in section 3.

15. The \mathbb{C} here is because we are looking at representations on complex vector spaces; if we were looking at representations on vector spaces over some field \mathbb{F} , we would take ${}^L G(\mathbb{F})$.

16. The appropriately modified Galois group is called the Weil-Deligne group.

1 What Are Combinatorial and Geometric Group Theory?

Groups and geometry are ubiquitous in mathematics, groups because the symmetries (or AUTOMORPHISMS [I.3 §4.1]) of any mathematical object in any context form a group and geometry because it allows one to think intuitively about abstract problems and to organize families of objects into spaces from which one may gain some global insight.

The purpose of this article is to introduce the reader to the study of infinite, discrete groups. I shall discuss both the combinatorial approach to the subject that held sway for much of the twentieth century and the more geometric perspective that has led to an enormous flowering of the subject in the last twenty years. I hope to convince the reader that the study of groups is a concern for all of mathematics rather than something that belongs particularly to the domain of algebra.

The principal focus of *geometric group theory* is the interaction of geometry/topology and group theory, through group actions and through suitable translations of geometric concepts into group theory. One wants to develop and exploit this interaction for the benefit of both geometry/topology and group theory. And, in keeping with our assertion that groups are important throughout mathematics, one hopes to illuminate and solve problems from elsewhere in mathematics by encoding them as problems in group theory.

Geometric group theory acquired a distinct identity in the late 1980s but many of its principal ideas have their roots in the end of the nineteenth century. At that time, low-dimensional topology and *combinatorial group theory* emerged entwined. Roughly speaking, combinatorial group theory is the study of groups defined in terms of *presentations*, that is, by means of generators and relations. In order to follow the rest of this introduction the reader must first understand what these terms mean. Since their definitions would require an unacceptably long break in the flow of our discussion, I will postpone them to the next section, but I strongly advise the reader who is unfamiliar with the meaning of the expression $\Gamma = \langle a_1, \dots, a_n \mid r_1, \dots, r_m \rangle$ to pause and read that section before continuing with this one.

The rough definition of combinatorial group theory just given misses the point that, like many parts of mathematics, it is a subject defined more by its core problems and its origins than by its fundamental definitions. The initial impetus for the subject came from

the description of discrete groups of hyperbolic isometries and, most particularly, the discovery of the FUNDAMENTAL GROUP [IV.6 §2] of a MANIFOLD [I.3 §6.9] by POINCARÉ [VI.61] in 1895. The group-theoretic issues that emerged were brought into sharp focus by the work of Tietze and Dehn in the first decade of the twentieth century and drove much of combinatorial group theory for the remainder of the century.

Not all of the epoch-defining problems came from topology: other areas of mathematics threw up fundamental questions as well. Here are some of the forms they took: Does there exist a group of the following type? Which groups have the following property? What are the subgroups of ...? Is the following group infinite? When can one determine the structure of a group from its finite quotients? In the sections that follow I shall attempt to illustrate the mathematical culture associated with questions of this kind, but let me immediately mention some easily stated but difficult classical problems. (i) Let G be a group that is finitely generated and suppose that there is some positive integer n such that $x^n = 1$ for every x in G . Must G be finite? (ii) Is there a finitely presented group Γ and a surjective homomorphism $\phi : \Gamma \rightarrow \Gamma$ such that $\phi(y) = 1$ for some $y \neq 1$? (iii) Does there exist a finitely presented, infinite, SIMPLE GROUP [I.3 §3.3]? (iv) Is every countable group isomorphic to a subgroup of a finitely generated group, or even a finitely presented group?

The first of these questions was asked by Burnside in 1902 and the second by Hopf in connection with his study of degree-1 maps between manifolds. I shall present the answers to all four questions (in section 5) to illustrate an important aspect of both combinatorial and geometric group theory: one develops techniques that allow the construction of *explicit groups* with prescribed properties. Such constructions are of particular interest when they illustrate the diversity of possible phenomena in other branches of mathematics.

Another kind of question that raises basic issues in combinatorial group theory takes the form: Does there exist an algorithm to determine whether or not a group (or given elements of a group) has such-and-such a property? For example, does there exist an algorithm that can take any finite presentation and decide in a finite number of steps whether or not the group presented is trivial? Questions of this type led to a profound and mutually beneficial interaction between group theory and logic, given full voice by the Higman embedding theorem, which we shall discuss in section 6. Moreover, via the conduit of combinatorial

PUP: proofreader asked for more specific cross-reference, but I think that was due to a misunderstanding that we're referring here to section 5 of this article. I mistakenly used the section symbol before. Now using the word, and in paragraph below, so OK now?

group theory, logic has influenced topology as well: one uses group-theoretic constructions to show, for example, that there is no algorithm to determine which pairs of compact triangulated manifolds are homeomorphic in dimensions 4 and above. This shows that certain kinds of classification results that have been obtained in two and three dimensions do not have higher-dimensional analogues.

One might reasonably regard combinatorial group theory as the attempt to develop algebraic techniques to solve the types of questions described above, and in the course of doing so to identify classes of groups that are worthy of particular study. This last point, the question of which groups deserve our attention, is tackled head-on in the final section of this article.

Some of the triumphs of combinatorial group theory are intrinsically combinatorial in nature, but many more have had their true nature revealed by the introduction of geometric techniques in the past twenty years. A fine example of this is the way in which Gromov's insights have connected algorithmic problems in group theory to so-called filling problems in Riemannian geometry. Moreover, the power of geometric group theory is by no means confined to improving the techniques of combinatorial group theory: it naturally leads one to think about many other issues of fundamental importance. For example, it provides a context in which one can illuminate and vastly extend classical RIGIDITY THEOREMS [V.26], such as that of Mostow. The key to applications such as this is the idea that finitely generated groups can usefully be regarded as geometric objects in their own right. This idea has its origins in the work of CAYLEY [VI.46] (1878) and Dehn (1905) but its full force was recognized and promoted by Gromov, starting in the 1980s. It is the key idea that underpins the later sections of this article.

2 Presenting Groups

How should one describe a group? An example will illustrate the standard way of doing so and give some idea of why it is often appropriate.

Consider the familiar tiling of the Euclidean plane by equilateral triangles. How might you describe the full group Γ_Δ of symmetries of this tiling, i.e., the rigid motions of the plane that send tiles to tiles? Let us focus on a single tile T and a particular edge e of T , and use this to pick out three symmetries. The first, which we shall call α , is the reflection of the plane in the line that contains e and the other two, β and γ , are the reflections in the lines that join the endpoints of e to the

midpoints of the opposite edges in T . With some effort one can convince oneself that every symmetry of the tiling can be obtained by performing these three operations repeatedly in a suitable order. One expresses this by saying that the set $\{\alpha, \beta, \gamma\}$ *generates* the group Γ_Δ .

A further useful observation is that if one performs the operation α twice, the tiling is returned to its original position: that is, $\alpha^2 = 1$. Likewise, $\beta^2 = \gamma^2 = 1$. One can also verify that $(\alpha\beta)^6 = (\alpha\gamma)^6 = (\beta\gamma)^3 = 1$.

It turns out that the group Γ_Δ is completely determined by these facts alone, a statement that we summarize by the notation

$$\Gamma_\Delta = \langle \alpha, \beta, \gamma \mid \alpha^2, \beta^2, \gamma^2, (\alpha\beta)^6, (\alpha\gamma)^6, (\beta\gamma)^3 \rangle.$$

The aim of the rest of this section is to say in more detail what this means.

To begin with, notice that from the facts we are given we can deduce others: for example, bearing in mind that $\beta^2 = \gamma^2 = (\beta\gamma)^3 = 1$, we can show that

$$(\gamma\beta)^3 = (\gamma\beta)^3(\beta\gamma)^3 = 1$$

as well (where the last equality follows after repeatedly canceling pairs of the form $\beta\beta$ or $\gamma\gamma$). We wish to convey the idea that in Γ_Δ there are no relationships between the generators except those that follow from the facts above by this kind of argument.

Now let us try to say this more formally. We define a *set of generators* for a group Γ to be a subset $S \subset \Gamma$ such that every element of Γ is equal to some product of elements of S and their inverses. That is, every element can be written in the form $s_1^{\varepsilon_1} s_2^{\varepsilon_2} \cdots s_n^{\varepsilon_n}$, where each s_i is an element of S and each ε_i is 1 or -1 . We then call a product of this kind a *relation* if it is equal to the identity in Γ .

There is an awkward ambiguity here. When we talk about “the product” of some elements of Γ , it sounds as though we are referring to another element of Γ , but we certainly did not mean this at the end of the last paragraph: a relation is not the identity element of Γ but rather a *string of symbols* such as $ab^{-1}a^{-1}bc$ that yields the identity in Γ when you interpret a , b , and c as generators in the set S . In order to be clear about this, it is useful to define another group, known as the *free group* $F(S)$.

For concreteness we shall describe the free group with three generators, taking our set S to be $\{a, b, c\}$. A typical element is a “word” in the elements of S and their inverses, such as the expression $ab^{-1}a^{-1}bc$ considered in the previous paragraph. However, we sometimes regard two words as the same: for instance,

$abcc^{-1}ac$ and $abab^{-1}bc$ are the same because they become identical when we cancel out the inverse pairs cc^{-1} and $b^{-1}b$. More formally, we define two such words to be *equivalent* and say that the elements of the free group are the EQUIVALENCE CLASSES [I.2 §2.3]. To multiply words together, we just concatenate them: for instance, the product of ab^{-1} and $bcca$ is $ab^{-1}bcca$, which we can shorten to $acca$. The identity is the “empty word.” This is the free group on three generators a, b , and c . It should be clear how to generalize it to an arbitrary set S , though we shall continue to discuss the set $S = \{a, b, c\}$.

A more abstract way of characterizing the free group on a, b , and c is to say that it has the following *universal property*: if G is any group and ϕ is any function from $S = \{a, b, c\}$ to G , then there is a unique homomorphism Φ from $F(S)$ to G that takes a to $\phi(a)$, b to $\phi(b)$, and c to $\phi(c)$. Indeed, if we want Φ to have these properties, then our definition is forced upon us: for example, $\Phi(ab^{-1}ca)$ will have to be $\phi(a)\phi(b)^{-1}\phi(c)\phi(a)$, by the definition of a homomorphism. So the uniqueness is obvious. The rough reason that this definition really does give rise to a well-defined homomorphism is that the only equations that are true in $F(S)$ are ones that are true in all groups: in order for Φ not to be a homomorphism, one would need a relation to hold in $F(S)$ that did not hold in G , but this is impossible.

Now let us return to our example Γ_Δ . We would like to prove that it is (isomorphic to) the “freest” group with generators α, β , and γ that satisfies the relations $\alpha^2 = \beta^2 = \gamma^2 = (\alpha\beta)^6 = (\alpha\gamma)^6 = (\beta\gamma)^3 = 1$. But what exactly is this “freest” group that we are claiming is isomorphic to Γ_Δ ?

To avoid confusion about the meaning of α, β , and γ (are they elements of Γ_Δ or of the group that we are trying to construct that will turn out to be isomorphic to Γ_Δ ?) we shall use the letters a, b , and c when we answer this question. Thus, we are trying to build the “freest” group with generators a, b , and c that satisfies the relations $a^2 = b^2 = c^2 = (ab)^6 = (ac)^6 = (bc)^3 = 1$, which we denote by $G = \langle a, b, c \mid a^2, b^2, c^2, (ab)^6, (ac)^6, (bc)^3 \rangle$.

There are two ways of going about this task. One is to imitate the above discussion of the free group itself, except that now we say that two words are equivalent if you can get from one to the other by inserting or deleting not just inverse pairs but also one of the words $a^2, b^2, c^2, (ab)^6, (ac)^6$, or $(bc)^3$. For example, ab^2c is equivalent to ac in this group. G is then defined to be

the set of equivalence classes of words with the product coming from concatenation.

A neater way to obtain G is more conceptual and exploits the universal property of the free group. As G is to be generated by a, b , and c , the universal property of the free group $F(S)$ tells us that there will have to be a unique homomorphism Φ from $F(S)$ to G such that $\Phi(a) = a$, $\Phi(b) = b$, and $\Phi(c) = c$. Moreover, we require that all of $a^2, b^2, c^2, (ab)^6, (ac)^6$, and $(bc)^3$ must map to the identity element in G . It follows that the KERNEL [I.3 §4.1] of Φ is a NORMAL SUBGROUP [I.3 §3.3] of $F(S)$ that contains the set $R = \{a^2, b^2, c^2, (ab)^6, (ac)^6, (bc)^3\}$. Let us write $\langle\langle R \rangle\rangle$ for the smallest normal subgroup of $F(S)$ that contains R (or equivalently the intersection of all normal subgroups of $F(S)$ that contain R). Then there is a surjective homomorphism from the QUOTIENT [I.3 §3.3] $F(S)/\langle\langle R \rangle\rangle$ to any group that is generated by a, b , and c and satisfies the relations $a^2 = b^2 = c^2 = (ab)^6 = (ac)^6 = (bc)^3 = 1$. This quotient itself is the group we are looking for: it is the largest group generated by a, b , and c that satisfies the relations in R .

Our assertion about Γ_Δ is that it is isomorphic to the group $G = \langle a, b, c \mid a^2, b^2, c^2, (ab)^6, (ac)^6, (bc)^3 \rangle$ that we have just described (in two ways). More precisely, the map from $F(S)/\langle\langle R \rangle\rangle$ to Γ_Δ that takes a to α , b to β , and c to γ is an isomorphism.

The above construction is very general. If we are given a group Γ , then a *presentation* of Γ is a set S that generates Γ , together with a set $R \subset F(S)$ of relations, such that Γ is isomorphic to the quotient $F(S)/\langle\langle R \rangle\rangle$. If both S and R are finite sets, one says that the presentation is finite. A group is *finitely presented* if it has a finite presentation.

We can also define presentations in the abstract, without mentioning a group Γ in advance: given any set S and any subset $R \subset F(S)$, we just define $\langle S \mid R \rangle$ to be the group $F(S)/\langle\langle R \rangle\rangle$. This is the “freest” group generated by S that satisfies the relations in R : the only relations that hold in $\langle S \mid R \rangle$ are the ones that can be deduced from the relations R .

A psychological advantage of switching to this more abstract setting is that, whereas previously we began with a group Γ and asked how we might present it, we can now write down group presentations at will, starting with any set S and prescribing a set of words R in the symbols $S^{\pm 1}$. This gives us a very flexible way of constructing a wide variety of groups. We might, for example, use a group presentation to encode a question from elsewhere in mathematics. We could then ask

PUP: Tim says that readers won't mind this seemingly tautological sentence.

about the properties of the group thus defined, and see what they had to tell us about our original problem.

3 Why Study Finitely Presented Groups?

Groups arise across the whole of mathematics as *groups of automorphisms*. These are maps from an object to itself that preserve all of the defining structure: two examples are the invertible LINEAR MAPS [I.3 §4.2] from a VECTOR SPACE [I.3 §2.3] to itself, and the homeomorphisms from a TOPOLOGICAL SPACE [III.92] to itself. Groups encapsulate the essence of symmetry and for this reason demand our attention. We are driven to understand their general nature, identify groups that deserve particular attention, and develop techniques for constructing new groups (from old ones, or from new ideas). And, reversing the process of abstraction, when *given* a group, we want to find concrete instances of it. For example, we might like to realize it as the group of automorphisms of some interesting object, with the aim of illuminating the nature of both the object and the group. (See the article on REPRESENTATION THEORY [IV.9] for more on this theme.)

3.1 Why Present Groups in Terms of Generators and Relations?

The short answer is that this is the form in which groups often “appear in nature.” This is particularly true in topology. Before looking at a general result that illustrates this point, let us examine a simple example. Consider the group D of all isometries of \mathbb{R} that are generated by the reflections at the points 0, 1, and 2: that is, the group generated by the three functions α_0 , α_1 , and α_2 , which take x to $-x$, $2-x$, and $4-x$, respectively. You may recognize this group to be the infinite dihedral group, and you may notice that the generator α_2 is superfluous, since it can be generated from α_0 and α_1 . But let us close our eyes to these observations as we let a presentation emerge from the action.

To this end, we choose an open interval U with the property that the images of U under the maps in D cover the whole of the real line, say $U = (-\frac{1}{2}, \frac{3}{2})$. Now let us record two pieces of data: the only elements of D (apart from the identity) that fail to move U completely off itself are α_0 and α_1 , and, among all products of length at most 3 in those two letters, the only nontrivial ones that act as the identity on \mathbb{R} are α_0^2 and α_1^2 . You may like to prove that $\langle \alpha_0, \alpha_1 \mid \alpha_0^2, \alpha_1^2 \rangle$ is a presentation of D .

This is in fact a special case of a general result, which we now state. (The proof of it is somewhat involved.) Let X be a topological space that is both PATH CONNECTED [IV.6 §1] and SIMPLY CONNECTED [III.95], and let Γ be a group of homeomorphisms from X to itself. Then any choice of path-connected open subset $U \subset X$ such that the images of U cover all of X gives rise to a presentation $\Gamma = \langle S \mid R \rangle$, where $S = \{ \gamma \in \Gamma \mid \gamma(U) \cap U \neq \emptyset \}$ and R consists of all words $w \in F(S)$ of length at most 3 such that $w = 1$ in Γ . Thus, the identification of a suitable subset U provides one with a presentation of Γ , and the task of a group theorist is to determine the nature of the group from this information.

To see how difficult this task is, you might like to consider the groups

$$G_n = \langle a_1, \dots, a_n \mid a_i^{-1} a_{i+1} a_i a_{i+1}^{-2}, i = 1, \dots, n \rangle,$$

where we interpret $i+1$ as 1 when $i = n$. One of G_3 and G_4 is trivial and the other is infinite. Can you decide which is which?

To illustrate a more subtle point, let us consider a finitely presented group that we perhaps feel we understand: the group Γ_Δ that we were discussing earlier. If we want to describe this group to a blind friend unfamiliar with the triangular tiling of the plane, what can we say to make her understand the group, or at least convince her that we understand the group?

Our friend might reasonably ask us to list the elements of our group, so we begin to describe them as products (words) in the given generators. But as we begin to do so we hit a problem: we do not want to list any element more than once and in order to avoid redundancy we have to know which pairs of words w_1, w_2 represent the same element of Γ_Δ ; equivalently, we must be able to recognize which words $w_1^{-1} w_2$ are relations in the group. Determining which words are relations is called the *word problem* for the group. Even in Γ_Δ this takes some work, and in the groups G_n we quickly find ourselves at a loss.

Note that as well as allowing one to list the elements of the group effectively, a solution to the word problem also allows one to determine the multiplication table, since deciding whether $w_1 w_2 = w_3$ is the same as deciding whether $w_1 w_2 w_3^{-1} = 1$.

3.2 Why Finitely Presented Groups?

The packaging of infinite objects into finite amounts of data arises throughout mathematics in the various guises of COMPACTNESS [III.9]. Finite presentation is basically a compactness condition: a group can be

finitely presented if and only if it is the fundamental group of a reasonable compact space, as we shall see later.

Another good reason for studying finitely presented groups is that the Higman embedding theorem (to be discussed later) allows us to encode questions about arbitrary TURING MACHINES [IV.20 §1.1] as questions about such groups and their subgroups.

4 The Fundamental Decision Problems

In exploring the geometry and topology of low-dimensional manifolds at the beginning of the twentieth century, Max Dehn saw that many of the problems that he was wrestling with could be “reduced” to questions about finitely presented groups. For example, he gave a simple formula for associating with a KNOT DIAGRAM [III.46] a finite presentation of a group. There was one relation for each crossing in the diagram and he argued that the resulting group would be isomorphic to \mathbb{Z} if and only if the knot was the unknot: that is, if and only if it could be continuously deformed into a circle. It is extremely hard to tell by staring at a knot diagram whether it is actually the unknot, so this seems like a useful reduction until one realizes that it can be just as hard to tell whether a finitely presented group is isomorphic to \mathbb{Z} . For example, here is the presentation of \mathbb{Z} that Dehn’s recipe associates with one of smallest possible pictures of the unknot, namely a diagram with just four crossings:

$$\langle a_1, a_2, a_3, a_4, a_5 \mid \\ a_1^{-1} a_3 a_4^{-1}, a_2 a_3^{-1} a_1, a_3 a_4^{-1} a_2^{-1}, a_4 a_5^{-1} a_4 a_3^{-1} \rangle.$$

Thus Dehn’s investigations led him to understand how difficult it is to extract information from a group presentation. In particular, he was the first to identify the fundamental role of the word problem, which we alluded to earlier, and he was one of the first to begin to understand that there are fundamental problems associated with the challenge of developing *algorithms* that extract knowledge from well-defined objects such as group presentations. In his famous article of 1912 Dehn writes:

The general discontinuous group is given by n generators and m relations between them. ... Here *there are above all three fundamental problems* whose solution is very difficult and which will not be possible without a penetrating study of the subject.

1. **The identity [word] problem:** An element of the group is given as a product of generators. One

is required to give a method whereby it may be decided in a finite number of steps whether this element is the identity or not.

2. **The transformation [conjugacy] problem:** Any two elements S and T of the group are given. A method is sought for deciding the question whether S and T can be transformed into each other, i.e., whether there is an element U of the group satisfying the relation

$$S = UTU^{-1}.$$

3. **The isomorphism problem:** Given two groups, one is to decide whether they are isomorphic or not (and further, whether a given correspondence between the generators of one group and elements of the other is an isomorphism or not).

We shall take these problems as the starting point for three lines of enquiry. First, we shall work toward an outline of the proof that all of these problems are, in a strict sense, unsolvable for general finitely presented groups.

The second use that we shall make of Dehn’s problems is to hold them up as fundamental measures of complexity for each of the classes of groups that we subsequently encounter. If we can prove, for example, that the isomorphism problem is solvable in one class of groups but not in another, then we will have given genuine substance to previously vague assertions to the effect that the second class is “harder.”

Finally, I want to make the point that geometry lies at the heart of the fundamental issues in combinatorial group theory: it may not be immediately obvious, but its implicit presence is nonetheless a fundamental trait of group theory and not something imposed for reasons of taste. To illustrate this point I shall explain how the study of the large-scale geometry of least-area disks in RIEMANNIAN MANIFOLDS [I.3 §6.10] is intimately connected with the study of the complexity of word problems in arbitrary finitely presented groups.

5 New Groups from Old

Suppose that you have two groups, G_1 and G_2 , and want to combine them to form a new group. The first method that is taught in a typical course on group theory is to take the Cartesian product $G_1 \times G_2$: a typical element has the form (g, h) with $g \in G_1$ and $h \in G_2$, and the product of (g, h) with (g', h') is defined to be (gg', hh') . The set of elements of the form (g, e) (where e is the identity of G_2) is a copy of G_1 inside $G_1 \times G_2$, and similarly the set of elements of the form (e, h) is a copy of G_2 .

These copies have nontrivial relations between their elements: for example, $(e, h)(g, e) = (g, e)(e, h)$. We would now like to take two groups Γ_1 and Γ_2 and combine them in a different way to form a group called the *free product* $\Gamma_1 * \Gamma_2$, which contains copies of Γ_1 and Γ_2 and as few additional relations as possible. That is, we would like there to be embeddings $i_j : \Gamma_j \hookrightarrow \Gamma_1 * \Gamma_2$ so that $i_1(\Gamma_1)$ and $i_2(\Gamma_2)$ generate $\Gamma_1 * \Gamma_2$ but they are not intertwined in any way. This requirement is neatly encapsulated by the following universal property: given any group G and any two homomorphisms $\phi_1 : \Gamma_1 \rightarrow G$ and $\phi_2 : \Gamma_2 \rightarrow G$, there should be a unique homomorphism $\Phi : \Gamma_1 * \Gamma_2 \rightarrow G$ such that $\Phi \circ i_j = \phi_j$ for $j = 1, 2$. (Less formally, Φ behaves like ϕ_1 on the copy of Γ_1 and behaves like ϕ_2 on the copy of Γ_2 .)

It is easy to check that this property characterizes $\Gamma_1 * \Gamma_2$ up to isomorphism, but it leaves open the question of whether $\Gamma_1 * \Gamma_2$ actually exists. (These are the standard pros and cons of defining an object by means of a universal property.) In the present setting, existence is easily established using presentations: let $\langle A_1 \mid R_1 \rangle$ be a presentation of Γ_1 and let $\langle A_2 \mid R_2 \rangle$ be a presentation of Γ_2 , with A_1 and A_2 disjoint, and then define $\Gamma_1 * \Gamma_2$ to be $\langle A_1 \sqcup A_2 \mid R_1 \sqcup R_2 \rangle$ (where \sqcup denotes a union of disjoint sets).

More intuitively, one can define $\Gamma_1 * \Gamma_2$ to be the set of alternating sequences $a_1 b_1 \cdots a_n b_n$ with each a_i belonging to Γ_1 and each b_j belonging to Γ_2 , with the extra condition that none of the a_i and b_j equals the identity, except possibly a_1 or b_n . The group operations in Γ_1 and Γ_2 extend to this set in an obvious way: for example, $(a_1 b_1 a_2)(a'_1 b'_1) = a_1 b_1 a'_2 b'_1$, where $a'_2 = a_2 a'_1$, except that if $a_2 a'_1 = 1$ then the product cancels down to $a_1 b'_2$, where $b'_2 = b_1 b'_1$.

Free products occur naturally in topology: if one has topological spaces X_1, X_2 with marked points $p_1 \in X_1$, $p_2 \in X_2$, then the FUNDAMENTAL GROUP [IV.6 §2] of the space $X_1 \vee X_2$ obtained from $X_1 \sqcup X_2$ by making the identification $p_1 = p_2$ is the free product of $\pi_1(X_1, p_1)$ and $\pi_1(X_2, p_2)$. The Seifert-van Kampen theorem tells one how to present the fundamental group of a space obtained by gluing X_1 and X_2 along larger subspaces. If the inclusion of the subspaces gives rise to an injection of fundamental groups, then one can express the fundamental group of the resulting space as an *amalgamated free product*, which we now define.

Let Γ_1 and Γ_2 be two groups. If some other group contains copies of Γ_1 and Γ_2 , then the intersection of those copies must contain the identity element. The

free product $\Gamma_1 * \Gamma_2$ was the freest group we could build that was subject to this minimal constraint. Now we shall insist that the copies of Γ_1 and Γ_2 intersect nontrivially, specify which of their subgroups must lie in the intersection, and build the freest group that satisfies this constraint.

Suppose, then, that A_1 is a subgroup of Γ_1 and that ϕ is an isomorphism from A_1 to a subgroup A_2 of Γ_2 . As in the example of the free product, one can define the “freest product that identifies A_1 and A_2 ” by means of a universal property. Again, one can establish the existence of such a group using presentations: if $\Gamma_1 = \langle S_1 \mid R_1 \rangle$ and $\Gamma_2 = \langle S_2 \mid R_2 \rangle$, the group we seek takes the form

$$\langle S_1 \sqcup S_2 \mid R_1 \sqcup R_2 \sqcup T \rangle.$$

Here, $T = \{u_a v_a^{-1} \mid a \in A_1\}$, where u_a is some word that represents a in (the presentation of) Γ_1 and v_a is a word that represents $\phi(a)$ in Γ_2 .

This group is called the *amalgamated free product of Γ_1 and Γ_2 along A_1 and A_2* . It is often described by the casual and ambiguous notation $\Gamma_1 *_{A_1=A_2} \Gamma_2$, or even $\Gamma_1 *_A \Gamma_2$, where $A \cong A_j$ is an abstract group.

Unlike with free products, it is no longer obvious that the maps $\Gamma_i \rightarrow \Gamma_1 *_A \Gamma_2$ implicit in this construction are injective, but they do turn out to be, as was shown by Schreier in 1927.

A related construction of Higman, Neumann, and Neumann in 1949 answers the following question: given a group Γ and an isomorphism $\psi : B_1 \rightarrow B_2$ between subgroups of Γ , can one always embed Γ in a bigger group so that ψ becomes the restriction to B_1 of a conjugation?

By now, having seen the idea in the context of both free products and amalgamated free products, the reader may guess how one goes about answering this question: one writes down the presentation of a universal candidate for the desired enveloping group, denoted $\Gamma *_{\psi}$, and then one sets about proving that the natural map from Γ to $\Gamma *_{\psi}$ (which takes each word to itself) is injective. Thus, given $\Gamma = \langle A \mid R \rangle$, we introduce a symbol $t \notin A$ (usually called the *stable letter*), we choose for each $b \in B_1$ words $\hat{b}, \tilde{b} \in F(A)$ with $\hat{b} = b$ and $\tilde{b} = \psi(b)$ in Γ , and we define

$$\Gamma *_{\psi} = \langle A, t \mid R, t \hat{b} t^{-1} \tilde{b}^{-1} \ (b \in B_1) \rangle.$$

This is the freest group we can build from Γ by adjoining a new element t and requiring it to satisfy all the equations we want it to, namely $t \hat{b} t^{-1} = \tilde{b}$ for every $b \in B_1$ (which we can think of as saying that $t b t^{-1} =$

PUP: thanks to the proofreader for spotting the typo in this display.

PUP: Tim tried rewriting this but couldn't find a better form of words. OK as it is?

$\psi(b)$). This group is called an *HNN extension* of Γ (after Higman, Neumann, and Neumann).

Now we must show that the natural map from Γ to $\Gamma *_{\phi}$ is injective. That is, if you take an element γ of Γ and regard it as an element of $\Gamma *_{\psi}$, you should not be able to use t and the relations in $\Gamma *_{\psi}$ to cancel γ down to the identity. This is proved with the help of the following more general result known as *Britton's lemma*. Suppose that w is a word in the free group $F(A, t)$. Then the only circumstances under which it can give rise to the identity in the group $\Gamma *_{\psi}$ are if either it does not involve t and represents the identity in Γ or it involves t but can be simplified in an obvious way by containing a "pinch." A pinch is a subword of the form tbt^{-1} , where b is a word in $F(A)$ that represents an element of B_1 (in which case we can replace it by $\psi(b)$), or one of the form $t^{-1}b't$, where b' represents an element of B_2 (in which case we can replace it by $\psi^{-1}(b')$). Thus, if you are given a word that involves t and contains no pinches, then you know that it cannot be canceled down to the identity.

A similar noncancellation result holds for the amalgamated free product $\Gamma_1 *_{A_1=A_2} \Gamma_2$. If g_1, \dots, g_n belong to Γ_1 but not to A_1 and h_1, \dots, h_n belong to Γ_2 but not to A_2 , then the word $g_1 h_1 g_2 h_2 \cdots g_n h_n$ cannot equal the identity in $\Gamma_1 *_{A_1=A_2} \Gamma_2$.

These noncancellation results do far more than show that the natural homomorphisms we have been considering are injective: they also demonstrate further aspects of freeness in amalgamated free products and HNN extensions. For example, suppose that in the amalgamated free product $\Gamma_1 *_{A_1=A_2} \Gamma_2$ we can find an element g of Γ_1 that generates an infinite group that intersects A_1 in the identity and an element h of Γ_2 that does the same for A_2 . Then the subgroup of $\Gamma_1 *_{A_1=A_2} \Gamma_2$ generated by g and h is the free group on those two generators. With a little more effort, one can deduce that any finite subgroup of $\Gamma_1 *_{A_1=A_2} \Gamma_2$ has to be conjugate to a subgroup of the obvious copy of either Γ_1 or Γ_2 . Similarly, the finite subgroups of $\Gamma *_{\psi}$ are conjugates of subgroups of Γ . We shall exploit these facts in the constructions that follow.

There are many ways of combining groups that I have not mentioned here. I have chosen to focus on amalgamated free products and HNN extensions partly because they lead to transparent solutions of the basic problems discussed below but more because of their primitive appeal and the way in which they arise naturally in the calculation of fundamental groups. They also mark the beginning of *arboreal group theory*,

which we will discuss later. If space allowed, I would go on to describe semidirect and wreath products, which are also indispensable tools of the group theorist.

Before turning to some applications of HNN extensions and amalgamated free products, I want to return to the Burnside problem, which asks if there exist finitely generated infinite groups all of whose elements have a given finite order. This question generated important developments throughout the twentieth century, particularly in Russia. It is appropriate to mention it here because it provides another illustration of the fact that it can be useful to study a universal object in order to solve a general question.

5.1 The Burnside Problem

Given an exponent m , one clarifies the problem at hand by considering the *free Burnside group* $B_{n,m}$ given by the presentation $\langle a_1, \dots, a_n \mid R_m \rangle$, where R_m consists of all m th powers in the free group $F(a_1, \dots, a_n)$. It is clear that $B_{n,m}$ maps onto any group with at most n generators in which every element has order dividing m . Therefore, there exists a finitely generated infinite group with all elements of the same finite order if and only if, for suitable values of n and m , the group $B_{n,m}$ is infinite. Thus, a question that takes the form, Does there exist a group such that ...?, becomes a question about just one group.

Novikov and Adian showed in 1968 that $B_{n,m}$ is infinite when $n \geq 2$ and $m \geq 667$ is odd. Determining the exact range of values for which $B_{n,m}$ is infinite is an active area of research. Of far greater interest is the open question of whether there exist finitely presented infinite groups that are quotients of $B_{n,m}$. Zelmanov was awarded the Fields Medal for proving that each $B_{n,m}$ has only finitely many finite quotients.

5.2 Every Countable Group Can Be Embedded in a Finitely Generated Group

Given a countable group G we list its elements, g_0, g_1, g_2, \dots , taking g_0 to be the identity. We then take a free product of G with an infinite cyclic group $\langle s \rangle \cong \mathbb{Z}$. Let Σ_1 be the set of all elements of $G * \mathbb{Z}$ of the form $s_n = g_n s^n$ with $n \geq 1$. Then the subgroup $\langle \Sigma_1 \rangle$ generated by Σ_1 is isomorphic to the free group $F(\Sigma_1)$. Similarly, if we let $\Sigma_2 = \{s_2, s_3, \dots\}$ (so it is Σ_1 with the element $s_1 = g_1 s$ removed), then $\langle \Sigma_2 \rangle$ is isomorphic to $F(\Sigma_2)$. It follows that the map $\psi(s_n) = s_{n+1}$ gives rise to an isomorphism from $\langle \Sigma_1 \rangle$ to $\langle \Sigma_2 \rangle$. Now take the HNN extension $(G * \mathbb{Z}) *_{\psi}$, whose stable letter we denote by

PUP: this technical term has to stay as it is.

PUP: 'of' is correct here.

t . This group contains a copy of G , as we noted before. Moreover, since we have ensured that $ts_nt^{-1} = s_{n+1}$ for every $n \geq 1$, it can be generated by just the three elements s_1, s , and t . Thus, we have embedded an arbitrary countable group into a group with three generators. (We leave the reader to think about how one can vary this construction to produce a group with two generators.)

5.3 There Are Uncountably Many Nonisomorphic Finitely Generated Groups

This was proved by B. H. Neumann in 1932. Since there are infinitely many primes, there are uncountably many nonisomorphic groups of the form $\bigoplus_{p \in P} \mathbb{Z}_p$, where P is an infinite set of primes. We have seen that each of these groups can be embedded in a finitely generated group, and our earlier comments on finite subgroups of HNN extensions show that no two of the resulting finitely generated groups are isomorphic.

5.4 An Answer to Hopf's Question

A group G is called *Hopfian* if every surjective homomorphism from G to G is an isomorphism. Most familiar groups have this property: for example, finite groups obviously do, as do \mathbb{Z}^n (as you can prove using linear algebra) and free groups. So too do groups of matrices such as $\mathrm{SL}_n(\mathbb{Z})$, as we shall discuss in a moment. An example of a non-Hopfian group is the group of all infinite sequences of integers (under pointwise addition), since the function that takes (a_1, a_2, a_3, \dots) to (a_2, a_3, a_4, \dots) is a surjective homomorphism that contains $(1, 0, 0, \dots)$ in its kernel. But is there a finitely presented example? The answer is yes, and Higman was the first to construct one. The following examples are due to Baumslag and Solitar.

Let $p \geq 2$ be an integer and identify \mathbb{Z} with the free group $\langle a \rangle$ generated by a single generator a . Then the subgroups $p\mathbb{Z}$ and $(p+1)\mathbb{Z}$ of \mathbb{Z} are identified with the powers of a^p and a^{p+1} , respectively. Let ψ be the isomorphism between these subgroups that takes a^p to a^{p+1} and consider the corresponding HNN extension B . This has presentation $B = \langle a, t \mid ta^{-p}t^{-1}a^{p+1} \rangle$. The homomorphism $\psi : B \rightarrow B$ defined by $t \mapsto t, a \mapsto a^p$ is clearly a surjection but its kernel contains, for example, the element $c = ata^{-1}t^{-1}a^{-2}tat^{-1}a$, which does not contain a pinch and is therefore not equal to the identity, by Britton's lemma. (If you want to convince yourself how useful this lemma is, set $p = 3$ and try to prove directly that c is not equal to the identity in the group B just defined.)

5.5 A Group that Has No Faithful Linear Representation

One can show that a finitely generated group G of matrices over any field is *residually finite*, which means that for each nontrivial element $g \in G$ there exists a finite group Q and a homomorphism $\pi : G \rightarrow Q$ with $\pi(g) \neq 1$. For example, if you are given an element $g \in \mathrm{SL}_n(\mathbb{Z})$, then you can pick an integer m bigger than the absolute values of all the entries in g (which is an $n \times n$ matrix) and consider the homomorphism from $\mathrm{SL}_n(\mathbb{Z})$ to $\mathrm{SL}_n(\mathbb{Z}/m\mathbb{Z})$ that reduces the matrix entries mod m . The image of g in the finite group $\mathrm{SL}_n(\mathbb{Z}/m\mathbb{Z})$ is clearly nontrivial.

Non-Hopfian groups are not residually finite, and hence are not isomorphic to a group of matrices over any field. One can see that the non-Hopfian group B defined above is not residually finite by considering what happens to the nontrivial element c . We saw that there was a surjective homomorphism $\psi : B \rightarrow B$ with $\psi(c) = 1$. Let c_n be an element such that $\psi^n(c_n) = c$ (which exists since ψ is a surjection). If there were a homomorphism π from B to a finite group Q with $\pi(c) \neq 1$, then we would have infinitely many distinct homomorphisms from B to Q , namely the compositions $\pi \circ \psi^n$; these are distinct because $\pi \circ \psi^m(c_n) = 1$ if $m > n$ and $\pi \circ \psi^n(c_n) = \pi(c) \neq 1$. This is a contradiction, since a homomorphism from a finitely generated group to a finite group is determined by what it does to the generators, so there can only be finitely many such homomorphisms.

5.6 Infinite Simple Groups

Britton's lemma actually tells us more than that $c \neq 1$: the subgroup Λ of B generated by t and c is in fact a free group on those generators. Thus we may form the amalgamated free product Γ of two copies of B , denoted B_1 and B_2 , by gluing together the two copies of Λ with the isomorphism $c_1 \mapsto t_2, t_1 \mapsto c_2$. We have seen that in any finite quotient of $\Gamma = B_1 *_\Lambda B_2$, the elements $c_1 (= t_2)$ and $c_2 (= t_1)$ must have trivial image, and it is easy to deduce from this that in fact the quotient must be trivial. Thus Γ is an infinite group with no finite quotients. It follows that the quotient of Γ by any maximal proper normal subgroup is also infinite (and it is simple by maximality).

The simple group that we have constructed is infinite and finitely generated but it is not finitely presentable. Finitely presented infinite simple groups do exist, but they are much harder to construct.

6 Higman's Theorem and Undecidability

We have seen that there are uncountably many (non-isomorphic) finitely generated groups. But as there are only countably many finitely *presented* groups, only countably many finitely generated groups can be subgroups of finitely presented groups. Which ones are they?

A complete answer to this question is provided by a beautiful and deep theorem proved by Graham Higman in 1961, which says, roughly, that the groups that arise are all those that are algorithmically describable. (If you have no idea what this means, even roughly, then you might like to read THE INSOLUBILITY OF THE HALTING PROBLEM [V.23] before continuing with this section.)

A set S of words over a finite alphabet A is called *recursively enumerable* if there is some algorithm (or more formally, Turing machine) that can produce a complete list of the elements of S . A case of particular interest is when A is just a singleton, in which case a word is determined by its length and we can think of S as a set of nonnegative integers. The elements of S need not be listed in a sensible order, so having an algorithm that produces an exhaustive list of S does not mean that one can use the algorithm to determine that some given word w does *not* belong to S : if you imagine standing by your computer as it enumerates S , there will not in general come a time when you can say to yourself, "If it was going to appear, then it would have done so by now," and therefore be certain that it is not in S . If you want an algorithm with this further property, then you need the stronger notion of a *recursive set*, which is a set S such that S and its complement are *both* recursively enumerable. Then you can list all the elements that belong to S and you can also list all the elements that do not belong to S .

A finitely generated group is said to be *recursively presentable* if it has a presentation with a finite number of generators and a recursively enumerable set of defining relations. In other words, such a group is not necessarily finitely presented, but at least the presentation of the group is "nice" in the sense that it can be generated by some algorithm.

Higman's embedding theorem states that a *finitely generated group G is recursively presentable if and only if it is isomorphic to a subgroup of a finitely presented group*.

To get a feeling for how nonobvious this is, you might consider the following presentation of the group of all rationals under addition, in which the generator a_n

corresponds to the fraction $1/n!$:

$$Q = \langle a_1, a_2, \dots \mid a_n^n = a_{n-1} \ \forall n \geq 2 \rangle.$$

Higman's theorem tells us that Q can be embedded in a finitely presented group, but no truly explicit embedding is known.

The power of Higman's theorem is illustrated by the ease with which it implies the celebrated undecidability results that were rightly regarded as watersheds of twentieth-century mathematics. In order to make this case convincingly, I shall give a complete proof (except that I shall assume some of the facts mentioned earlier) that there exist finitely presented groups with unsolvable word problems, and also that there are sequences of finitely presented groups among which one cannot decide isomorphism. We shall also see how these group-theoretic results can be used to translate undecidability phenomena into topology.

The basic seed of undecidability comes from the fact that there are recursively enumerable subsets $S \subset \mathbb{N}$ that are not recursive. Using this fact one can readily construct finitely generated groups with an unsolvable word problem: given such a set of integers S we consider

$$J = \langle a, b, t \mid t(b^n a b^{-n})t^{-1} = b^n a b^{-n} \ \forall n \in S \rangle.$$

This is the HNN extension of the free group $F(a, b)$ associated with the identity map $L \rightarrow L$, where L is the subgroup generated by $\{b^n a b^{-n} : n \in S\}$. Britton's lemma tells us that the word $w_m = t(b^m a b^{-m})t^{-1}(b^m a^{-1} b^{-m})$ equals 1 in J if and only if $m \in S$, and by definition there is no algorithm to decide if $m \in S$, so we cannot decide which of the w_m are relations. Thus J has an unsolvable word problem.

That there exist finitely presented groups for which the word problem is unsolvable is a much deeper fact, but with Higman's embedding theorem at hand the proof becomes almost trivial: Higman tells us that J can be embedded in a finitely presented group Γ , and it is a relatively straightforward exercise to show that if one cannot decide which words in the generators of J represent the identity, then one cannot decide for arbitrary words in the generators of Γ either.

Once one has a finitely presented group with an unsolvable word problem, it is easy to translate undecidability into all manner of other problems. For example, suppose that $\Gamma = \langle A \mid R \rangle$ is a finitely presented group with an unsolvable word problem, where $A = \{a_1, \dots, a_n\}$ and no a_i equals the identity in Γ . For each word w made out of the letters in A and their inverses,

define a group Γ_w to have presentation

$$\langle A, s, t \mid R, t^{-1}(s^i a_i s^{-i}) t (s^i w s^{-i}), i = 1, \dots, n \rangle.$$

It is not hard to show that if $w = 1$ in Γ then Γ_w is the free group generated by s and t . If $w \neq 1$, then Γ_w is an HNN extension. In particular, it contains a copy of Γ , and hence has an unsolvable word problem, which means that it cannot be a free group. Thus, since there is no algorithm to decide whether $w = 1$ in Γ , one cannot decide which of the groups Γ_w are isomorphic to which others.

A variant of this argument shows that there is no algorithm to determine whether or not a given finitely presented group is trivial.

We shall see in a moment that every finitely presented group G is the fundamental group of some compact four-dimensional manifold. By following a standard proof of this theorem with considerable care, Markov proved in 1958 that in dimensions 4 and above there is no algorithm to decide which compact manifolds (presented as simplicial complexes, for example) are homeomorphic. His basic idea was to show that if there were an algorithm to determine which triangulated 4-manifolds are homeomorphic, then one could use it to determine which finitely presented groups are trivial, which we know is impossible. In order to implement this idea one has to be careful to arrange that the 4-manifolds associated with different presentations of the trivial group are homeomorphic: this is the delicate part of the argument.

Strikingly, there does exist an algorithm to decide which compact three-dimensional manifolds are isomorphic. This is an extremely deep theorem that relies in particular on Perelman's solution to THURSTON'S GEOMETRIZATION CONJECTURE [IV.7 §2.4].

7 Topological Group Theory

Let us change perspective now and look at the symbols $P \equiv \langle a_1, \dots, a_2 \mid r_1, \dots, r_m \rangle$ through the eyes of a topologist. Instead of interpreting P as a recipe for constructing a group, we regard it as a recipe for constructing a TOPOLOGICAL SPACE [III.92], or more specifically a *two-dimensional complex*. Such spaces consist of points, called *vertices*, some of which are linked by directed paths, called *edges*, or 1-cells. If a collection of such 1-cells forms a cycle, then it can be filled in with a *face*, or 2-cell: topologically speaking, each face is a disk with a directed cycle as its boundary.

To see what this complex is, let us first consider the standard presentation $P \equiv \langle a, b \mid aba^{-1}b^{-1} \rangle$ of \mathbb{Z}^2 .

(This is generated by a and b and the relation tells us that $ab = ba$.) We begin with a graph K^1 that has a single vertex and two edges (which are loops) that are directed and labeled a and b . Next, we take a square $[0, 1] \times [0, 1]$, the sides of which are directed and labeled a, b, a^{-1}, b^{-1} as we proceed around the boundary. Imagine gluing the boundary of the square to the graph so as to respect the labeling of edges: with a bit of thought, you should be able to see that the result is a torus, that is, a surface in the shape of a bagel. An observation that turns out to be important is that the fundamental group of the torus is \mathbb{Z}^2 , the group we started with.

The idea of “gluing” is made precise by the use of *attaching maps*: we take a continuous map ϕ from the boundary of the square S to the graph K^1 that sends the corners of the square to the vertex of K^1 and sends each side (minus its vertices) homeomorphically onto an open edge. The torus is then the quotient of $K^1 \sqcup S$ by the equivalence relation that identifies each x in the boundary of the square with its image $\phi(x)$.

With this more abstract language in hand, it is easy to see how the above construction generalizes to arbitrary presentations: given a presentation $P \equiv \langle a_1, \dots, a_n \mid r_1, \dots, r_m \rangle$, one takes a graph with a single vertex and n oriented loops, which are labeled a_1, \dots, a_n . Then for each r_j one attaches a polygonal disk by gluing its boundary circuit to the sequence of oriented edges that traces out the word r_j .

In general, the result will not be a surface as it was for $\langle a, b \mid aba^{-1}b^{-1} \rangle$. Rather, it will be a two-dimensional complex with singularities along the edges and at the vertex. You may find it instructive to do some more examples. From $\langle a \mid a^2 \rangle$ one gets the projective plane; from $\langle a, b, c, d \mid aba^{-1}b^{-1}, cdc^{-1}d \rangle$ one gets a torus and a Klein bottle stuck together at a point. Picturing the 2-complex for $\langle a, b \mid a^2, b^3, (ab)^3 \rangle$ is already rather difficult.

The construction of $K(P)$ is the beginning of *topological group theory*. The Seifert-van Kampen theorem (mentioned earlier) implies that the fundamental group of $K(P)$ is the group presented by P . But the group no longer sits inertly in the form of an inscrutable presentation—now it acts on the UNIVERSAL COVERING [III.95] of $K(P)$ by homeomorphisms known as “deck transformations.” Thus, through the simple construction of $K(P)$ (and the elegant theory of covering spaces in topology) we achieve our aim of realizing an abstract finitely presented group as the group of symmetries of an object with a potentially rich structure, on which we

PUP: ‘ $cdc^{-1}d$ ’ is correct here.

can bring global geometric and topological techniques to bear.

To obtain an improved topological model for our group, we can embed $K(P)$ in \mathbb{R}^5 (just as one can embed a finite GRAPH [III.34] in \mathbb{R}^3) and consider the compact four-dimensional manifold M obtained by taking all points that are a small fixed distance from the image. (I am assuming that the embedding is suitably “tame,” which one can arrange.) The mental picture to strive for here is a higher-dimensional analogue of the surface (sleeve) one gets by taking the points in \mathbb{R}^3 that are a small fixed distance from an embedded graph. The fundamental group of M is again the group presented by P , so now we have our arbitrary finitely presented group acting on a manifold (the universal cover of M). This allows us to use the tools of analysis and differential geometry.

The constructions of $K(P)$ and M establish the more difficult implication of the theorem, promised earlier, that a group can be finitely presented if and only if it is the fundamental group of a compact cell complex and of a compact 4-manifold. This result raises several natural questions. First, are there better, more informative, topological models for an arbitrary finitely presented group Γ ? And if not, then what can one say about the classes of groups defined by the natural constraints that arise when one tries to improve the model? For example, we would like to construct a lower-dimensional manifold with fundamental group Γ , enabling us to exploit our physical insight into three-dimensional geometry. But it turns out that the fundamental groups of compact three-dimensional manifolds are very special; this observation lies near the heart of a great deal of mathematics at the end of the twentieth century. Other interesting fields open up when one asks which groups arise as the fundamental groups of compact spaces satisfying CURVATURE [III.13] conditions, or constraints coming from complex geometry.

A particularly rich set of constraints comes from the following question. Can one arrange for an arbitrary finitely presented group to be the fundamental group of a compact space (a complex or manifold, perhaps) whose universal cover is CONTRACTIBLE [IV.6 §2]? This is a natural question from the point of view of topology because a space with a contractible universal cover is, up to HOMOTOPY [IV.6 §2], completely determined by its fundamental group. If the fundamental group is Γ , then such a space is called a *classifying space* for Γ and its homotopy-invariant properties provide a rich array

of invariants for the group Γ (getting away from the gross dependence that $K(P)$ has on P rather than Γ).

If our earlier discussion of how hard it is to recognize Γ from P has left you very skeptical about whether this dependence can actually be removed, then your skepticism is well-founded: there are many obstructions to the construction of compact classifying spaces for an arbitrary finitely presented group; the study of them (under the generic name *finiteness conditions*) is a rich area at the interface of modern group theory, topology, and homological algebra.

One aspect of this area is the search for natural conditions that ensure the *existence* of compact classifying spaces (not necessarily manifolds). This is one of several places where manifestations of nonpositive curvature play a fundamental role in modern group theory. More combinatorial conditions also arise. For example, Lyndon proved that for any presentation $P \equiv \langle A \mid r \rangle$ where the single defining relation $r \in F(A)$ is not a nontrivial power, the universal cover of $K(P)$ is contractible.

A neighboring and highly active area of research concerns questions of uniqueness and rigidity for classifying spaces. (Here, as is common, the word *rigidity* is used to describe a situation in which requiring two objects to be equivalent in an apparently weak sense forces them to be equivalent in an apparently stronger sense.) For example, the (open) *Borel conjecture* asserts that if two compact manifolds have isomorphic fundamental groups and contractible universal covers, then those manifolds must be homeomorphic.

I have been talking mostly about realizing groups as fundamental groups, which led to certain free actions. That is, we could interpret the elements of the group as symmetries of a topological space and none of these symmetries had any fixed points. Before moving on to geometric group theory I should point out that there are many situations in which the most illuminating actions of a group are not free: one instead allows well-understood stabilizers. (The *stabilizer* of a point is the set of all symmetries in the group that leave that point fixed.) For example, the natural way in which to study Γ_Δ is by its action on the triangulated plane, each vertex of which is left unmoved by twelve symmetries.

A deeper illustration of the merits of seeking insight into algebraic structure through nonfree actions on suitable topological spaces comes from the Bass-Serre theory of groups acting on trees, which subsumes the theory of amalgamated free products and HNN extensions, whose potency we saw earlier. (This theory and

its extensions often go under the heading of *arboreal group theory*.)

A *tree* is a connected graph that has no circuits in it. It is helpful to regard it as a METRIC SPACE [III.58] in which each edge has length 1. The group actions that one allows on trees are those that take edges to edges isometrically, never flipping an edge.

If a group Γ acts on a set X (in other words, if it can be regarded as a group of symmetries of X), then the *orbit* of a point $x \in X$ is the set of all its images gx with $g \in \Gamma$. A group Γ can be expressed as an amalgamated free product $A *_C B$ if and only if it acts on a tree in such a way that there are two orbits of vertices, one orbit of edges, and stabilizers A, B, C (where A and B are the stabilizers of adjacent vertices and intersect in C , which is the edge stabilizer). HNN extensions correspond to actions with one orbit of vertices and one orbit of edges. Thus, amalgamated free products and HNN extensions appear as *graphs of groups*, which are the basic objects of Bass–Serre theory. These objects allow one to recover groups acting on trees from the quotient data of the action, i.e., the quotient space (which is a graph) and the pattern of edge and vertex stabilizers.

An early benefit of Bass–Serre theory is a transparent and instructive proof that any finite subgroup of $A *_C B$ is conjugate to a subgroup of either A or B : given any set V of vertices in a tree, there is a unique vertex or mid-point x minimizing $\max\{d(x, v) \mid v \in V\}$; one applies this observation with V an orbit of the finite subgroup; x provides a fixed point for the action of the subgroup; and any point stabilizer is conjugate to a subgroup of either A or B .

Arboreal group theory goes much deeper than this first application suggests. It is the basis for a decomposition theory of finitely presented groups from which it emerges, for example, that there is an essentially canonical maximal splitting of an arbitrary finitely presented group as a graph of groups with cyclic edge stabilizers. This provides a striking parallel with the decomposition theory of 3-manifolds, a parallel that extends far beyond a mere analogy and accounts for much of the deepest work in geometric group theory in the past ten years. If you want to learn more about this, search the literature for *JSJ decompositions*. You may also want to search for *complexes of groups*, which provide the appropriate higher-dimensional analogue for graphs of groups.

8 Geometric Group Theory

Let us refresh the image of $K(P)$ in our mind’s eye by thinking again about the presentation $P \equiv \langle a, b \mid aba^{-1}b^{-1} \rangle$ of \mathbb{Z} . The complex $K(P)$, as we saw earlier, is a torus. Now the torus can be defined as the quotient of the Euclidean plane \mathbb{R}^2 by the action of the group \mathbb{Z}^2 (where the point $(m, n) \in \mathbb{Z}^2$ acts as the translation $(x, y) \mapsto (x + m, y + n)$): in fact, \mathbb{R}^2 , with an appropriate square tiling, is the universal cover of the torus. If we look at the orbit of the point 0 under this action, it forms a copy of \mathbb{Z}^2 , and one can thereby see the large-scale geometry of \mathbb{Z}^2 laid out for us. We can make the idea of the “geometry of \mathbb{Z}^2 ” precise by decreeing that edges of the tiling have length 1 and defining the *graph distance* between vertices to be the length of the shortest path of edges connecting them.

As this example shows, the construction of $K(P)$ involves the two main (intertwined) strands of geometric group theory. In the first and more classical strand, one studies actions of groups on metric and topological spaces in order to elucidate the structures of both the space and the group (as with the action of \mathbb{Z}^2 on the plane in our example, or the action of the fundamental group of $K(P)$ on its universal cover in general). The quality of the insights that one obtains varies according to whether the action has or does not have certain desirable properties. The action of \mathbb{Z}^2 on \mathbb{R}^2 consists of isometries on a space with a fine geometric structure, and the quotient (the torus) is compact. Such actions are in many ways ideal, but sometimes one accepts weaker admission criteria in order to obtain a more diverse class of groups, and sometimes one demands even more structure in order to narrow the focus and study groups and spaces of an exceptional, but for that reason interesting, character.

This first strand of geometric group theory mingles with the second. In the second strand, one regards finitely generated groups as geometric objects in their own right equipped with *word metrics*, which are defined as follows. Given a finite generating set S for a group Γ , one defines the *Cayley graph* of Γ by joining each element $y \in \Gamma$ by an edge to each element of the form ys or ys^{-1} with $s \in S$ (which is the same as the graph formed by the edges of the universal covering of $K(P)$). The distance $d_S(y_1, y_2)$ between y_1 and y_2 is then the length of the shortest path from y_1 to y_2 if all edges have length 1. Equivalently, it is the length of the shortest word in the free group on S that is equal to $y_1^{-1}y_2$ in Γ .

The word metric and the Cayley graph depend on the choice of generating set but their large-scale geometry does not. In order to make this idea precise, we introduce the notion of a *quasi-isometry*. This is an equivalence relation that identifies spaces that are similar on a large scale. If X and Y are two metric spaces, then a quasi-isometry from X to Y is a function $\phi : X \rightarrow Y$ with the following two properties. First, there are positive constants c , C , and ϵ such that $cd(x, x') - \epsilon \leq d(\phi(x), \phi(x')) \leq Cd(x, x') + \epsilon$: this says that ϕ distorts sufficiently large distances by at most a constant factor. Second, there is a constant C' such that for every $y \in Y$ there is some $x \in X$ for which $d(\phi(x), y) \leq C'$: this says that ϕ is a “quasi-surjection” in the sense that every element of Y is close to the image of an element of X .

Consider for example the two spaces \mathbb{R}^2 and \mathbb{Z}^2 , where the metric on \mathbb{Z}^2 is given by the graph distance defined earlier. In this case the map $\phi : \mathbb{R}^2 \rightarrow \mathbb{Z}^2$ that takes (x, y) to $(\lfloor x \rfloor, \lfloor y \rfloor)$ (where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x) is easily seen to be a quasi-isometry: if the Euclidean distance d between two points (x, y) and (x', y') is at least 10, say, then the graph distance between $(\lfloor x \rfloor, \lfloor y \rfloor)$ and $(\lfloor x' \rfloor, \lfloor y' \rfloor)$ will certainly lie between $\frac{1}{2}d$ and $2d$. Notice how little we care about the local structure of the two spaces: the map ϕ is a quasi-isometry despite not even being continuous.

It is not hard to check that if ϕ is a quasi-isometry from X to Y , then there is a quasi-isometry ψ from Y to X that “quasi-inverts” ϕ , in the sense that every x in X is at most a bounded distance from $\psi\phi(x)$ and every y in Y is at most a bounded distance from $\phi\psi(y)$. Once one has established this, it is easy to see that quasi-isometry is an equivalence relation.

Returning to Cayley graphs and word metrics, it turns out that if you take two different sets of generators for the same group, then the resulting Cayley graphs will be quasi-isometric. Thus, any property of a Cayley graph that is invariant under quasi-isometry will be a property not just of the graph but of the group itself. When dealing with such invariants we are free to think of Γ itself as a space (since we do not care which Cayley graph we form), and we can replace it by any metric space that is quasi-isometric to it, such as the universal cover of a closed Riemannian manifold with fundamental group Γ (whose existence we discussed earlier). Then the tools of analysis can be brought to bear on it.

A fundamental fact, discovered independently by many people and often called the *Milnor–Švarc lemma*,

provides a crucial link between the two main strands of geometric group theory. Let us call a metric space X a *length space* if the distance between each pair of points is the infimum of the lengths of paths joining them. The Milnor–Švarc lemma states that if a group Γ acts nontrivially as a set of isometries of a length space X , and if the quotient is compact, then Γ is finitely generated and quasi-isometric to X (for any choice of word metric).

We have seen an example of this already: \mathbb{Z}^2 is quasi-isometric to the Euclidean plane. Less obviously, the same is true of Γ_Δ . (Consider the map that takes each element α of Γ_Δ to the point of \mathbb{Z}^2 nearest $\alpha(0)$.)

The fundamental group of a compact Riemannian manifold is quasi-isometric to the universal cover of that manifold. Therefore, from the point of view of quasi-isometry invariants, the study of such manifolds is equivalent to the study of arbitrary finitely presented groups. In a moment we will discuss some nontrivial consequences of this equivalence. But first let us reflect on the fact that, when finitely generated groups are considered as metric objects in the framework of large-scale geometry, they present us with a new challenge: we should *classify finitely generated groups up to quasi-isometry*.

This is an impossible task, of course, but nevertheless serves as a beacon in modern geometric group theory, one that has guided us toward many beautiful theorems, particularly under the general heading of rigidity. For example, suppose that you come across a finitely generated group Γ that is reminiscent of \mathbb{Z}^n on a large scale: in other words, quasi-isometric to it. We are not necessarily given any algebraically defined map between this mystery group and \mathbb{Z}^n , and yet it transpires that such a group must contain a copy of \mathbb{Z}^n as a subgroup of finite index.

At the heart of this result is *Gromov’s polynomial growth theorem*, a landmark theorem published in 1981. This theorem concerns the number of points within a distance r of the identity in a finitely generated group Γ . This will be a function $f(r)$, and Gromov was interested in how the function $f(r)$ grows as r tends to infinity, and what that tells us about the group Γ .

If Γ is an Abelian group with d generators, then it is not hard to see that $f(r)$ is at most $(2r + 1)^d$ (since each generator is raised to a power between $-r$ and r). Thus, in this case $f(r)$ is bounded above by a polynomial in r . At the other extreme, if Γ is a free group with two generators a and b , say, then $f(r)$ is exponentially large, since all sequences of length r that consist of a s

and bs (and not their inverses) give different elements of Γ .

Given this sharp contrast in behavior, one might wonder whether requiring $f(r)$ to be bounded above by a polynomial forces Γ to exhibit a great deal of commutativity. Fortunately, there is a much-studied definition that makes this idea precise. Given any group G and any subgroup H of G , the *commutator* $[G, H]$ is the subgroup generated by all elements of the form $ghg^{-1}h^{-1}$, where g belongs to G and h belongs to H . If G is Abelian, then $[G, H]$ contains just the identity. If G is not Abelian, then $[G, G]$ forms a group G_1 that contains other elements besides the identity, but it may be that $[G, G_1]$ is trivial. In that case, one says that G is a two-step nilpotent group. In general, a *k-step nilpotent* group G is one where, if you form a sequence by setting $G_0 = G$ and $G_{i+1} = [G, G_i]$ for each i , then you eventually reach the trivial group, and the first time you do so is at G_k . A *nilpotent* group is a group that is k -step nilpotent for some k .

Gromov's theorem states that a group has polynomial growth if and only if it has a nilpotent subgroup of finite index. This is a quite extraordinary fact: the polynomial growth condition is easily seen to be independent of the choice of word metric and to be an invariant of quasi-isometry. Thus the seemingly rigid and purely algebraic condition of having a nilpotent subgroup of finite index is in fact a quasi-isometry invariant, and therefore a flabby, robust characteristic of the group.

In the past fifteen years quasi-isometric rigidity theorems have been established for many other classes of groups, including lattices in semisimple Lie groups and the fundamental groups of compact 3-manifolds (where the classification up to quasi-isometry involves more than algebraic equivalences), as well as various classes defined in terms of their graph of group decompositions. In order to prove theorems of this type, one must identify nontrivial invariants of quasi-isometry that allow one to distinguish and relate various classes of spaces. In many cases such invariants come from the development of suitable analogues of the tools of algebraic topology, modified so that they behave well with respect to quasi-isometries rather than continuous maps.

9 The Geometry of the Word Problem

It is time to explain the comments I made earlier about the geometry inherent in the basic decision problems of combinatorial group theory. I shall concentrate exclusively on the geometry of the word problem.

Gromov's *filling theorem* describes a startlingly intimate connection between the highly geometric study of disks with minimal area in RIEMANNIAN GEOMETRY [I.3 §6.10] and the study of word problems, which seems to belong more to algebra and logic.

On the geometric side, the basic object of study is the *isoperimetric function* $\text{Fill}_M(l)$ of a smooth compact manifold M . Given any closed path of length l , there is a disk of minimal area that is bounded by that path. The largest such area, over all closed paths of length l , is defined to be $\text{Fill}_M(l)$. Thus, the isoperimetric function is the smallest function of which it is true to say that every closed path of length l can be filled by a disk of area at most $\text{Fill}_M(l)$.

The image to have in mind here is that of a soap film: if one twists a circular wire of length l in Euclidean space and dips it in soap, the film that forms has area at most $l^2/4\pi$, whereas if one performs the same experiment in HYPERBOLIC SPACE [I.3 §6.6], the area of the film is bounded by a linear function of l . Correspondingly, the isoperimetric functions of \mathbb{E}^n and \mathbb{H}^n (and quotients of them by groups of isometries) are quadratic and linear, respectively. In a moment we shall discuss what types of isoperimetric functions arise when one considers other geometries (more precisely, compact Riemannian manifolds).

To state the filling theorem we need to think about the algebraic side as well. Here, we identify a function that measures the complexity of a direct attack on the word problem for an arbitrary finitely presented group $\Gamma = \langle A \mid R \rangle$. If we wish to know whether a word w equals the identity in Γ and do not have any further insight into the nature of Γ , then there is not much we can do other than repeatedly insert or remove the given relations $r \in R$.

Consider the simple example $\Gamma = \langle a, b \mid b^2a, baba \rangle$. In this group aba^2b represents the identity. How do we prove this? Well,

$$\begin{aligned} aba^2b &= a(b^2a)ba^2b = ab(baba)ab \\ &= abab = a(baba)a^{-1} = aa^{-1} = 1. \end{aligned}$$

Now let us think about the proof geometrically, via the Cayley graph. Since $aba^2b = 1$ in the group Γ , we obtain a cycle in this graph if we start at the identity and go along edges labeled a, b, a, a, b , in that order (in which case we visit the vertices $1, a, ab, aba, aba^2, aba^2b = 1$). The equalities in the proof can be thought of as a way of "contracting" this cycle down to the identity by means of inserting or deleting small loops: for instance, we could insert b, a, b, a into the list of edge

PUP: 'flabby' means 'not brittle' here and this combination of words is OK.

directions, since *baba* is a relation, or we could delete a trivial loop of the form a, a^{-1} . This contraction can be given a more topological character if we turn our Cayley graph into a two-dimensional complex by filling in each small loop with a *face*. Then the contraction of the original cycle consists in gradually moving it across these small faces.

Thus, the difficulty of demonstrating that a word w equals the identity is intimately connected with the *area* of w , denoted $\text{Area}(w)$, which can be thought of algebraically as the smallest sequence of relations you need to insert and delete to turn w into the identity, or geometrically as the smallest number of faces you need to make a disk that fills the cycle represented by w .

The *Dehn function* $\delta_\Gamma : \mathbb{N} \rightarrow \mathbb{N}$ bounds $\text{Area}(w)$ in terms of the length $|w|$ of the word w : $\delta_\Gamma(n)$ is the largest area of any word of length at most n that equals 1 in Γ . If the Dehn function grows rapidly, then the word problem is hard, since there are short words that are equal to the identity, but their area is very large, so that any demonstration that they are equal to the identity has to be very long. Results bounding the Dehn function are called *isoperimetric inequalities*.

The subscript on δ_Γ is somewhat misleading since different finite presentations of the same group will in general yield different Dehn functions. This ambiguity is tolerated because it is tightly controlled: if the groups defined by two finite presentations are isomorphic, or just quasi-isometric, then the corresponding Dehn functions have similar growth rates. More precisely, they are *equivalent*, with respect to what is sometimes called the *standard equivalence relation* “ \simeq ” of geometric group theory: given two monotone functions $f, g : [0, \infty) \rightarrow [0, \infty)$, one writes $f \preceq g$ if there exists a constant $C > 0$ such that $f(l) \leq Cg(Cl + C) + Cl + C$ for all $l \geq 0$, and $f \simeq g$ if $f \preceq g$ and $g \preceq f$; and one extends this relation to include functions from \mathbb{N} to $[0, \infty)$.

You will have noticed a resemblance between the definitions of $\text{Fill}_M(l)$ and $\delta_\Gamma(n)$. The filling theorem relates them precisely: it states that *if M is a smooth compact manifold, then $\text{Fill}_M(l) \simeq \delta_\Gamma(l)$, where Γ is the fundamental group $\pi_1 M$ of M .*

For example, since \mathbb{Z}^2 is the fundamental group of the torus $T = \mathbb{R}^2/\mathbb{Z}^2$, which has Euclidean geometry, $\delta_{\mathbb{Z}^2}(l)$ is quadratic.

9.1 What Are the Dehn Functions?

We have seen that the complexity of word problems is related to the study of isoperimetric problems in

Riemannian and combinatorial geometry. Such insights have, in the last fifteen years, led to great advances in the understanding of the nature of Dehn functions. For example, one can ask for which numbers ρ the function n^ρ is a Dehn function. The set of all such numbers, which can be shown to be countable, is known as the *isoperimetric spectrum*, denoted IP , and it is now largely understood.

Following work by many authors, Brady and Bridson proved that the closure of IP is $\{1\} \cup [2, \infty)$. The finer structure of IP was described by Birget, Rips, and Sapir in terms of the time functions of Turing machines. A further result by the same authors and Ol’shanskii explains how fundamental Dehn functions are to understanding the complexity of arbitrary approaches to the word problem for finitely generated groups Γ : the word problem for Γ lies in NP if and only if Γ is a subgroup of a finitely presented group with polynomial Dehn function. (Here, NP is the class of problems in the famous “P versus NP” question: see COMPUTATIONAL COMPLEXITY [IV.20 §3] for a description of this class.)

The structure of IP raises an obvious question: What can one say about the two classes of groups singled out as special—those with linear Dehn functions and those with quadratic ones? The true nature of the class of groups with a quadratic Dehn function remains obscure for the moment but there is a beautifully definitive description of those with a linear Dehn function: they are the *word hyperbolic groups*, which we shall discuss in the next section.

Not all Dehn functions are of the form n^α : there are Dehn functions such as $n^\alpha \log n$, for example, and others that grow more quickly than any iterated exponential, for example that of

$$\langle a, b \mid aba^{-1}bab^{-1}a^{-1}b^{-2} \rangle.$$

If Γ has unsolvable word problem, then $\delta_\Gamma(n)$ will grow faster than any recursive function (indeed this serves as a definition of such groups).

9.2 The Word Problem and Geodesics

A *closed geodesic* on a Riemannian manifold is a loop that locally minimizes distance, such as a loop formed by an elastic band when released on a perfectly smooth surface. Examples such as the great circles on a sphere or the waist of an hourglass show that manifolds may contain closed geodesics that are *null-homotopic*: that is, they can be moved continuously until they are reduced to a point. But can one construct a compact

topological manifold with the property that no matter what metric one puts on it there will always be infinitely many such geodesics? (Technically, if you go around a geodesic loop n times, then you get a geodesic; we avoid this by counting only “primitive” geodesics.)

From a purely geometric point of view this is a daunting problem: all specific metric information has been stripped away and one has to deal with an arbitrary metric on the floppy topological object left behind. But group theory provides a solution: *if the Dehn function of the fundamental group $\pi_1 M$ grows at least as fast as 2^{2^n} , then in any Riemannian metric on M there will be infinitely many closed geodesics that are null-homotopic.* The proof of this is too technical to sketch here.

10 Which Groups Should One Study?

Several special classes of groups have emerged from our previous discussion, such as nilpotent groups, 3-manifold groups, groups with linear Dehn functions, and groups with a single defining relation. Now we shall change viewpoint and ask which groups present themselves for study as we set out to explore the universe of all finitely presented groups, starting with the easiest ones.

The trivial group comes first, of course, followed by the finite groups. Finite groups are discussed in various other places in this volume, so I shall ignore them in what follows and adopt the approach of large-scale geometry, blurring the distinction between groups that have a common subgroup of finite index.

The first infinite group is surely \mathbb{Z} , but what comes next is open to debate. If one wants to retain the safety of commutativity, then finitely generated Abelian groups come next. Then, as one slowly relinquishes commutativity and control over growth and constructibility, one passes through the progressively larger classes of nilpotent, polycyclic, solvable, and elementary amenable groups. We have already met nilpotent groups in our discussion of Gromov’s polynomial-growth theorem. They crop up in many contexts as the most natural generalization of Abelian groups and much is known about them, not least because one can prove a great deal by induction on the k for which they are k -step nilpotent. One can also exploit the fact that G is built from the finitely generated Abelian groups G_i/G_{i+1} in a very controlled way. The larger class of polycyclic groups is built in a similar way, while finitely generated solvable groups are built in a finite number of steps from Abelian groups that need not be finitely

generated. This last class is not only larger but wilder; the isomorphism problem is solvable among polycyclic groups, for example, but unsolvable among solvable groups. By definition a group G is solvable if its *derived series*, defined inductively by $G^{(n)} = [G^{(n-1)}, G^{(n-1)}]$ with $G^{(0)} = G$, terminates in a finite number of steps.

PUP: this is fine.

The concept known as *amenability* forms an important link between geometry, analysis, and group theory. Solvable groups are amenable but not vice versa. It is not quite the case that a finitely presented group is amenable if and only if it does not contain a free subgroup of rank 2, but for a novice this serves as a good rule of thumb.

Now, let us return to \mathbb{Z} in a more adventurous frame of mind, throw away the security of commutativity, and start taking free products instead. In this more liberated approach, finitely generated free groups appear after \mathbb{Z} as the first groups in the universe. What comes next? Thinking geometrically, we might note that free groups are precisely those groups that have a tree as a Cayley graph and then ask which groups have Cayley graphs that are *tree-like*.

A key property of a tree is that all of its triangles are degenerate: if you take any three points in the tree and join them by shortest paths, then every point in one of these paths is contained in at least one other path as well. This is a manifestation of the fact that trees are spaces of infinite negative curvature. To get a feeling for why, consider what happens when one rescales the metric on a space of bounded negative curvature such as the hyperbolic plane \mathbb{H}^2 . If we replace the standard distance function $d(x, y)$ by $(1/n)d(x, y)$ and let n tend to ∞ , then the curvature of this space (in the classical sense of differential geometry) tends to $-\infty$. This is captured by the fact that triangles look increasingly degenerate: there is a constant $\delta(n)$, with $\delta(n) \rightarrow 0$ as $n \rightarrow \infty$, such that any side of a triangle in the scaled hyperbolic space $(\mathbb{H}^2, (1/n)d)$ is contained in the $\delta(n)$ -neighborhood of the union of the other two sides. More colloquially, triangles in \mathbb{H}^2 are *uniformly thin* and get increasingly thin as one rescales the metric.

With this picture in mind, one might move a little away from trees by asking which groups have Cayley graphs in which all triangles are uniformly thin. (It makes little sense to specify the thinness constant δ since it will change when one changes generating set.) The answer is Gromov’s *hyperbolic groups*. This is a fascinating class of groups that has many equivalent definitions and arises in many contexts. For example, we have already met it as the class of groups that have

linear Dehn functions. (It is not at all obvious that these two definitions are equivalent.)

Gromov's great insight is that because the thin-triangles condition encapsulates so much of the essence of the large-scale geometry of negatively curved manifolds, hyperbolic groups share many of the rich properties enjoyed by the groups that act nicely by isometries on such spaces. Thus, for example, hyperbolic groups have only finitely many conjugacy classes of finite subgroups, contain no copy of \mathbb{Z}^2 , and (after accounting for torsion) have compact classifying spaces. Their conjugacy problems can be solved in less than quadratic time, and Sela showed that one can even solve the isomorphism problem among torsion-free hyperbolic groups. In addition to their many fascinating properties and natural definition, a further source of interest in hyperbolic groups is the fact that in a precise statistical sense, a *random finitely presented group* will be hyperbolic.

Spaces of negative and nonpositive curvature have played a central role in many branches of mathematics in the last twenty years. There is no room even to begin to justify this assertion here but it does guide us in where to look for natural enlargements of the class of hyperbolic groups: we want *nonpositively curved groups*, defined by requiring that their Cayley graphs enjoy a key geometric feature that cocompact groups of isometries inherit from simply connected spaces of nonpositive curvature ("CAT(0) spaces"). But in contrast to the hyperbolic case, the class of groups that one obtains varies considerably when one perturbs the definition, and delineating the resulting classes and their (rich) properties has been the subject of much research.

The added complications that one encounters when one moves from negative to nonpositive curvature are exemplified by the fact that the isomorphism problem is unsolvable in one of the most prominent classes that arises: the so-called *combable groups*.

Let us now return to free groups and ask which hyperbolic groups are the *immediate* neighbors of free groups. Remarkably, this vague question has a convincing answer.

One of the great triumphs of arboreal group theory is the proof that there is a finite description of the set $\text{Hom}(G, F)$ of homomorphisms from an arbitrary finitely generated group G to a free group F . The basic building blocks in this description are what Sela calls *limit groups*. One of the many ways of defining a limit group L is that for each finite subset $X \subset L$ there

should exist a homomorphism to a finitely generated free group that is injective on X .

Limit groups can also be defined as those whose FIRST-ORDER LOGIC [IV.23 §1] resembles that of a free group in a precise sense. To see how first-order logic can be used to say something nontrivial about a group, consider the sentence

$$\forall x, y, z$$

$$(xy \neq yx) \vee (yz \neq zy) \vee (xz = zx) \vee (y = 1).$$

A group with this property is *commutative transitive*: if x commutes with $y \neq 1$, and y commutes with z , then x commutes with z . Free groups and Abelian groups have this property but a direct product of non-Abelian free groups, for example, does not.

It is a simple exercise to show that free Abelian groups are limit groups. But if one restricts attention to groups that have precisely the same first-order logic as free groups, one gets a smaller class consisting only of hyperbolic groups. The groups in this class are the subject of intense scrutiny at the moment. They all have negatively curved two-dimensional classifying spaces, built from graphs and hyperbolic surfaces in a hierarchical manner. The fundamental groups Σ_g of closed surfaces of genus $g \geq 2$ lie in this class, lending substance to the traditional opinion in combinatorial group theory that, among nonfree groups, it is the groups Σ_g that resemble free groups F_n most closely.

Incorporating this opinion into our earlier discussion, we arrive at the view that the groups \mathbb{Z}^n , the free groups F_n , and the groups Σ_g are the most basic of infinite groups. This is the start of a rich vein of ideas involving the automorphisms of these groups. In particular, there are many striking parallels between their outer automorphism groups $\text{GL}_n(\mathbb{Z})$, $\text{Out}(F_n)$, and $\text{Mod}_g \cong \text{Out}(\Sigma_g)$ (the mapping class group). These three classes of groups play a fundamental role across a broad spectrum of mathematics. I have mentioned them here in order to make the point that, beyond the search for knowledge about natural classes of groups, there are certain "gems" in group theory that merit a deep and penetrating study in their own right. Other groups that people might suggest for this category include Coxeter groups (generalized reflection groups, for which I_Δ is a prototype) and Artin groups (particularly BRAID GROUPS [III.4], which again crop up in many branches of mathematics).

I have thrown classes of groups at you thick and fast in this last section. Even so, there are many fascinating classes of groups and important issues that

I have ignored completely. But so it must be, for as Higman's theorem assures us, the challenges, joys, and frustrations of finitely presented groups can never be exhausted.

Further Reading

- Bridson, M. R., and A. Haefliger. 1999. *Metric Spaces of Non-Positive Curvature*. Grundlehren der Mathematischen Wissenschaften, volume 319. Berlin: Springer.
- Gromov, M. 1984. Infinite groups as geometric objects. In *Proceedings of the International Congress of Mathematicians, Warszawa, Poland, 1983*, volume 1, pp. 385–92. Warsaw: PWN.
- . 1993. Asymptotic invariants of infinite groups. In *Geometric Group Theory*, volume 2. London Mathematical Society Lecture Note Series, volume 182. Cambridge: Cambridge University Press.
- Lyndon, R. C., and P. E. Schupp. 2001. *Combinatorial Group Theory*. Classics in Mathematics. Berlin: Springer.

IV.11 Harmonic Analysis

Terence Tao

1 Introduction

Much of analysis tends to revolve around the study of general classes of FUNCTIONS [I.2 §2.2] and OPERATORS [III.52]. The functions are often real-valued or complex-valued, but may take values in other sets, such as a VECTOR SPACE [I.3 §2.3] or a MANIFOLD [I.3 §6.9]. An operator is itself a function, but at a “second level,” because its domain and range are themselves spaces of functions: that is, an operator takes a function (or perhaps more than one function) as its input and returns a transformed function as its output. Harmonic analysis focuses in particular on the *quantitative* properties of such functions, and how these quantitative properties change when various operators are applied to them.¹

What is a “quantitative property” of a function? Here are two important examples. First, a function is said to be *uniformly bounded* if there is some real number M such that $|f(x)| \leq M$ for every x . It can often be useful to know that two functions f and g are “uniformly close,” which means that their difference $f - g$

is uniformly bounded with a small bound M . Second, a function is called *square integrable* if the integral $\int |f(x)|^2 dx$ is finite. The square integrable functions are important because they can be analyzed using the theory of HILBERT SPACES [III.37].

A typical question in harmonic analysis might then be the following: if a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is square integrable, its gradient ∇f exists, and all the n components of ∇f are also square integrable, does this imply that f is uniformly bounded? (The answer is yes when $n = 1$, and no, but only just, when $n = 2$; this is a special case of the *Sobolev embedding theorem*, which is of fundamental importance in the analysis of PARTIAL DIFFERENTIAL EQUATIONS [IV.12].) If so, what are the precise bounds one can obtain? That is, given the integrals of $|f|^2$ and $|(\nabla f)_i|^2$, what can you say about the uniform bound M that you obtain for f ?

Real and complex functions are of course very familiar in mathematics, and one meets them in high school. In many cases one deals primarily with SPECIAL FUNCTIONS [III.87]: polynomials, exponentials, trigonometric functions, and other very concrete and explicitly defined functions. Such functions typically have a very rich algebraic and geometric structure, and many questions about them can be answered exactly using techniques from algebra and geometry.

However, in many mathematical contexts one has to deal with functions that are not given by an explicit formula. For example, the solutions to ordinary and partial differential equations often cannot be given in an explicit algebraic form (as a composition of familiar functions such as polynomials, EXPONENTIAL FUNCTIONS [III.25], and TRIGONOMETRIC FUNCTIONS [III.94]). In such cases, how does one think about a function? The answer is to focus on its *properties* and see what can be deduced from them: even if the solution of a differential equation cannot be described by a useful formula, one may well be able to establish certain basic facts about it and be able to derive interesting consequences from those facts. Some examples of properties that one might look at are measurability, boundedness, continuity, differentiability, smoothness, analyticity, integrability, or quick decay at infinity. One is thus led to consider interesting *general classes* of functions: to form such a class one chooses a property and takes the set of all functions with that property. Generally speaking, analysis is much more concerned with these general classes of functions than with individual functions. (See also FUNCTION SPACES [III.29].)

1. Strictly speaking, this sentence describes the field of *real-variable harmonic analysis*. There is another field called *abstract harmonic analysis*, which is primarily concerned with how real- or complex-valued functions (often on very general domains) can be studied using symmetries such as translations or rotations (for instance, via the Fourier transform and its relatives); this field is of course related to real-variable harmonic analysis, but is perhaps closer in spirit to representation theory and functional analysis, and will not be discussed here.

This approach can in fact be useful even when one is analyzing a single function that is very structured and has an explicit formula. It is not always easy, or even possible, to exploit this structure and formula in a purely algebraic manner, and then one must rely (at least in part) on more analytical tools instead. A typical example is the *Airy function*

$$\text{Ai}(x) = \int_{-\infty}^{\infty} e^{i(x\xi + \xi^3)} d\xi.$$

Although this is defined explicitly as a certain integral, if one wants to answer such basic questions as whether $\text{Ai}(x)$ is always a convergent integral, and whether this integral goes to zero as $x \rightarrow \pm\infty$, it is easiest to proceed using the tools of harmonic analysis. In this case, one can use a technique known as the *principle of stationary phase* to answer both these questions affirmatively, although there is the rather surprising fact that the Airy function decays almost exponentially fast as $x \rightarrow +\infty$, but only polynomially fast as $x \rightarrow -\infty$.

Harmonic analysis, as a subfield of analysis, is particularly concerned not just with qualitative properties like the ones mentioned earlier, but also with *quantitative bounds* that relate to those properties. For instance, instead of merely knowing that a function f is bounded, one may wish to know *how* bounded it is. That is, what is the *smallest* $M \geq 0$ such that $|f(x)| \leq M$ for all (or almost all) $x \in \mathbb{R}$; this number is known as the *sup norm* or L^∞ -norm of f , and is denoted $\|f\|_{L^\infty}$. Or instead of assuming that f is square integrable one can quantify this by introducing the L^2 -norm $\|f\|_{L^2} = (\int |f(x)|^2 dx)^{1/2}$; more generally one can quantify p th-power integrability for $0 < p < \infty$ via the L^p -norm $\|f\|_{L^p} = (\int |f(x)|^p dx)^{1/p}$. Similarly, most of the other qualitative properties mentioned above can be quantified by a variety of NORMS [III.64], which assign a nonnegative number (or $+\infty$) to any given function and which provide some quantitative measure of one characteristic of that function. Besides being of importance in pure harmonic analysis, quantitative estimates involving these norms are also useful in applied mathematics, for instance in performing an error analysis of some numerical algorithm.

Functions tend to have infinitely many degrees of freedom, and it is thus unsurprising that the number of norms one can place on a function is infinite as well: there are many ways of quantifying how “large” a function is. These norms can often differ quite dramatically from each other. For instance, if a function f is very large for just a few values, so that its graph has tall, thin “spikes,” then it will have a very large L^∞ -norm,

but $\int |f(x)| dx$, its L^1 -norm, may well be quite small. Conversely, if f has a very broad and spread-out graph, then it is possible for $\int |f(x)| dx$ to be very large even if $|f(x)|$ is small for every x : such a function has a large L^1 -norm but a small L^∞ -norm. Similar examples can be constructed to show that the L^2 -norm sometimes behaves very differently from either the L^1 -norm or the L^∞ -norm. However, it turns out that the L^2 -norm lies “between” these two norms, in the sense that if one controls *both* the L^1 -norm *and* the L^∞ -norm, then one also automatically controls the L^2 -norm. Intuitively, the reason is that if the L^∞ -norm is not too large then one eliminates all the spiky functions, and if the L^1 -norm is small then one eliminates most of the broad functions; the remaining functions end up being well-behaved in the intermediate L^2 -norm. More quantitatively, we have the inequality

$$\|f\|_{L^2} \leq \|f\|_{L^1}^{1/2} \|f\|_{L^\infty}^{1/2},$$

which follows easily from the trivial algebraic fact that if $|f(x)| \leq M$, then $|f(x)|^2 \leq M|f(x)|$. This inequality is a special case of HÖLDER’S INEQUALITY [V.22], which is one of the fundamental inequalities in harmonic analysis. The idea that control of two “extreme” norms automatically implies further control on “intermediate” norms can be generalized tremendously and leads to very powerful and convenient methods known as *interpolation*, which is another basic tool in this area.

The study of a single function and all its norms eventually gets somewhat tiresome, though. Nearly all fields of mathematics become a lot more interesting when one considers not just objects, but also *maps* between objects. In our case, the objects in question are functions, and, as was mentioned in the introduction, a map that takes functions to functions is usually referred to as an *operator*. (In some contexts it is also called a TRANSFORM [III.93].) Operators may seem like fairly complicated mathematical objects—their inputs and outputs are functions, which in turn have inputs and outputs that are usually numbers—but they are in fact a very natural concept since there are many situations where one wants to transform functions. For example, *differentiation* can be thought of as an operator, which takes a function f to its derivative df/dx . This operator has a well-known (partial) inverse, *integration*, which takes f to the function F that is defined by the formula

$$F(x) = \int_{-\infty}^x f(y) dy.$$

A less intuitive, but particularly important, example is THE FOURIER TRANSFORM [III.27]. This takes f to a function \hat{f} , given by the formula

$$\hat{f}(x) = \int_{-\infty}^{\infty} e^{-2\pi i x y} f(y) dy.$$

It is also of interest to consider operators that take two or more inputs. Two particularly common examples are the *pointwise product* and *convolution*. If f and g are two functions, then their pointwise product fg is defined in the obvious way:

$$(fg)(x) = f(x)g(x).$$

The convolution, denoted $f * g$, is defined as follows:

$$f * g(x) = \int_{-\infty}^{\infty} f(y)g(x - y) dy.$$

This is just a very small sample of interesting operators that one might look at. The original purpose of harmonic analysis was to understand the operators that were connected to Fourier analysis, real analysis, and complex analysis. Nowadays, however, the subject has grown considerably, and the methods of harmonic analysis have been brought to bear on a much broader set of operators. For example, they have been particularly fruitful in understanding the solutions of various linear and nonlinear partial differential equations, since the solution of any such equation can be viewed as an operator applied to the initial conditions. They are also very useful in analytic and combinatorial number theory, when one is faced with understanding the oscillation present in various expressions such as exponential sums. Harmonic analysis has also been applied to analyze operators that arise in geometric measure theory, probability theory, ergodic theory, numerical analysis, and differential geometry.

A primary concern of harmonic analysis is to obtain both qualitative and quantitative information about the effects of these operators on generic functions. A typical example of a quantitative estimate is the inequality

$$\|f * g\|_{L^\infty} \leq \|f\|_{L^2} \|g\|_{L^2},$$

which is true for all $f, g \in L^2$. This result, which is a special case of *Young's inequality*, is easy to prove: one just writes out the definition of $f * g(x)$ and applies the CAUCHY-SCHWARZ INEQUALITY [V.22]. As a consequence, one can draw the qualitative conclusion that the convolution of two functions in L_2 is always continuous. Let us briefly sketch the argument, since it is an instructive one.

A fundamental fact about functions in L^2 is that any such function f can be approximated arbitrarily well

(in the L^2 -norm) by a function \tilde{f} that is continuous and *compactly supported*. (The second condition means that \tilde{f} takes the value zero everywhere outside some interval $[-M, M]$.) Given any two functions f and g in L^2 , let \tilde{f} and \tilde{g} be approximations of this kind. It is an exercise in real analysis to prove that $\tilde{f} * \tilde{g}$ is continuous, and it follows easily from the inequality above that $\tilde{f} * \tilde{g}$ is close to $f * g$ in the L^∞ -norm, since

$$f * g - \tilde{f} * \tilde{g} = f * (g - \tilde{g}) + (f - \tilde{f}) * \tilde{g}.$$

Therefore, $f * g$ can be approximated arbitrarily well in the L^∞ -norm by continuous functions. A standard result in basic real analysis (that a uniform limit of continuous functions is continuous) now tells us that $f * g$ is continuous.

Notice the general structure of this argument, which occurs frequently in harmonic analysis. First, one identifies a “simple” class of functions for which one can easily prove the result one wants. Next, one proves that every function in a much wider class can be approximated in a suitable sense by simple functions. Finally, one uses this information to deduce that the result holds for functions in the wider class as well. In our case, the simple functions were the continuous functions of finite support, the wider class consisted of square-integrable functions, and the suitable sense of approximation was closeness in the L^2 -norm.

We shall give some further examples of qualitative and quantitative analysis of operators in the next section.

2 Example: Fourier Summation

To illustrate the interplay between quantitative and qualitative results, we shall now sketch some of the basic theory of summation of Fourier series, which historically was one of the main motivations for studying harmonic analysis.

In this section, we shall consider functions f that are *periodic* with period 2π : that is, functions such that $f(x + 2\pi) = f(x)$ for all x . An example of such a function is $f(x) = 3 + \sin(x) - 2\cos(3x)$. A function like this, which can be written as a finite linear combination of functions of the form $\sin(nx)$ and $\cos(nx)$, is called a *trigonometric polynomial*. The word “polynomial” is used here because any such function can be expressed as a polynomial in $\sin(x)$ and $\cos(x)$, or alternatively, and somewhat more conveniently, as a polynomial in e^{ix} and e^{-ix} . That is, it can be written as $\sum_{n=-N}^N c_n e^{inx}$ for some N and some choice of coefficients ($c_n : -N \leq n \leq N$). If we know that f can

be expressed in this form, then we can work out the coefficient c_n quite easily: it is given by the formula

$$c_n = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-inx} dx.$$

It is a remarkable and very important fact that we can say something similar about a much wider class of functions—if, that is, we now allow *infinite* linear combinations. Suppose that f is a periodic function that is also continuous (or, more generally, that f is *absolutely integrable*, meaning that the integral of $|f(x)|$ between 0 and 2π is finite). We can then define the *Fourier coefficients* $\hat{f}(n)$ of f , using exactly the formula we had above for c_n :

$$\hat{f}(n) = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-inx} dx.$$

The example of trigonometric polynomials now suggests that one should have the identity

$$f(x) = \sum_{n=-\infty}^{\infty} \hat{f}(n) e^{inx},$$

expressing f as a sort of “infinite trigonometric polynomial,” but this is not always true, and even when it is true it takes some effort to justify it rigorously, or even to say precisely what the infinite sum means.

To make the question more precise, let us introduce for each natural number N the *Dirichlet summation operator* S_N . This takes a function f to the function $S_N f$ that is defined by the formula

$$S_N f(x) = \sum_{n=-N}^N \hat{f}(n) e^{inx}.$$

The question we would like to answer is whether $S_N f$ converges to f as $N \rightarrow \infty$. The answer turns out to be surprisingly complicated: not only does it depend on the assumptions that one places on the function f , but it also depends critically on how one defines “convergence.” For example, if we assume that f is continuous and ask for the convergence to be uniform, then the answer is very definitely no: there are examples of continuous functions f for which $S_N f$ does not even converge pointwise to f . However, if we ask for a weaker form of convergence, the answer is yes: $S_N f$ will necessarily converge to f in the L^p topology for any $0 < p < \infty$, and even though it does not have to converge pointwise, it will converge *almost everywhere*, meaning that the set of x for which $S_N f(x)$ does not converge to $f(x)$ has MEASURE [III.57] zero. If instead one assumes only that f is absolutely integrable, then it is possible for the partial sums $S_N f$ to diverge at every single point x , as well as being divergent in the

L^p topology for every p such that $0 < p \leq \infty$. The proofs of all of these results ultimately rely on very quantitative results in harmonic analysis, and in particular on various L^p -type estimates on the Dirichlet sum $S_N f(x)$, as well as estimates connected with the closely related *maximal operator*, which takes f to the function $\sup_{N>0} |S_N f(x)|$.

As these results are a little tricky to prove, let us first discuss a simpler result, in which the Dirichlet summation operators S_N are replaced by the *Fejér summation operators* F_N . For each N , the operator F_N is the average of the first N Dirichlet operators: that is, it is given by the formula

$$F_N = \frac{1}{N} (S_0 + \cdots + S_{N-1}).$$

It is not hard to show that if $S_N f$ converges to f , then so does $F_N f$. However, by averaging the $S_N f$ we allow cancellations to take place that sometimes make it possible for $F_N f$ to converge to f even when $S_N f$ does not. Indeed, here is a sketch of a proof that $F_N f$ converges to f whenever f is continuous and periodic—which, as we have seen, is far from true of $S_N f$.

In its basic structure, the argument is similar to the one we used when showing that the convolution of two functions in L^2 is continuous. Note first that the result is easy to prove when f is a trigonometric polynomial, since then $S_N f = f$ for every N from some point onward. Now the *Weierstrass approximation theorem* says that every continuous periodic function f can be uniformly approximated by trigonometric polynomials: that is, for every $\varepsilon > 0$ there is a trigonometric polynomial such that $\|f - g\|_{L^\infty} \leq \varepsilon$. We know that $F_N g$ is close to g for large N (since g is a trigonometric polynomial), and would like to deduce the same for f .

The first step is to use some routine trigonometric manipulation to prove the identity

$$F_N f(x) = \int_{-\pi}^{\pi} \frac{\sin^2(\frac{1}{2}N\gamma)}{N \sin^2(\frac{1}{2}\gamma)} f(x - \gamma) d\gamma.$$

The precise form of this expression is less important than two properties of the function

$$u(\gamma) = \frac{\sin^2(\frac{1}{2}N\gamma)}{N \sin^2(\frac{1}{2}\gamma)}$$

that we shall use. One is that $u(\gamma)$ is always nonnegative and the other is that $\int_{-\pi}^{\pi} u(\gamma) d\gamma = 1$. These two facts allow us to say that

$$\begin{aligned} F_N h(x) &= \int_{-\pi}^{\pi} u(\gamma) h(x - \gamma) d\gamma \\ &\leq \|h\|_{L^\infty} \int_{-\pi}^{\pi} u(\gamma) d\gamma = \|h\|_{L^\infty}. \end{aligned}$$

That is, $\|F_N h\|_{L^\infty} \leq \|h\|_{L^\infty}$ for any bounded function h .

To apply this result, we choose a trigonometric polynomial g such that $\|f - g\|_{L^\infty} \leq \varepsilon$ and let $h = f - g$. Then we find that $\|F_N h\|_{L^\infty} = \|F_N f - F_N g\|_{L^\infty} \leq \varepsilon$ as well. As mentioned above, if we choose N large enough, then $\|F_N g - g\|_{L^\infty} \leq \varepsilon$, and then we use the TRIANGLE INEQUALITY [V.22] to say that

$$\begin{aligned} \|F_N f - f\|_{L^\infty} \\ \leq \|F_N f - F_N g\|_{L^\infty} + \|F_N g - g\|_{L^\infty} + \|g - f\|_{L^\infty}. \end{aligned}$$

Since each term on the right-hand side is at most ε , this shows that $\|F_N f - f\|_{L^\infty}$ is at most 3ε . And since ε can be made arbitrarily small, this shows that $F_N f$ converges to f .

A similar argument (using MINKOWSKI'S INTEGRAL INEQUALITY [V.22] instead of the triangle inequality) shows that $\|F_N f\|_{L^p} \leq \|f\|_{L^p}$ for all $1 \leq p \leq \infty$, $f \in L^p$, and $N \geq 1$. As a consequence, one can modify the above argument to show that $F_N f$ converges to f in the L^p topology for every $f \in L^p$. A slightly more difficult result (relying on a basic result in harmonic analysis known as the *Hardy-Littlewood maximal inequality*) asserts that, for every $1 < p \leq \infty$, there exists a constant C_p such that one has the inequality $\|\sup_N |F_N f|\|_{L^p} \leq C_p \|f\|_{L^p}$ for all $f \in L^p$; as a consequence, one can show that $F_N f$ converges to f almost everywhere for every $f \in L^p$ and $1 < p \leq \infty$. A slight modification of this argument also allows one to treat the endpoint case when f is merely assumed to be absolutely integrable; see the discussion on the Hardy-Littlewood maximal inequality at the end of this article.

Now let us return briefly to Dirichlet summation. Using fairly sophisticated techniques in harmonic analysis (such as Calderón-Zygmund theory) one can show that when $1 < p < \infty$ the Dirichlet operators S_N are bounded in L^p uniformly in N . In other words, for every p in this range there exists a positive real number C_p such that $\|S_N f\|_{L^p} \leq C_p \|f\|_{L^p}$ for every function f in L^p and every nonnegative integer N . As a consequence, one can show that $S_N f$ converges to f in the L^p topology for all f in L^p and every p such that $1 < p < \infty$. However, the quantitative estimate on S_N fails at the endpoints $p = 1$ and $p = \infty$, and from this one can also show that the convergence result also fails at these endpoints (either by explicitly constructing a counterexample or by using general results such as the so-called *uniform boundedness principle*).

What happens if we ask for $S_N f$ to converge to f almost everywhere? Almost-everywhere convergence

does not follow from convergence in L^p when $p < \infty$, so we cannot use the above results to prove it. It turns out to be a much harder question, and was a famous open problem, eventually answered by CARLESON'S THEOREM [V.5] and an extension of it by Hunt. Carleson proved that one has an estimate of the form $\|\sup_N |S_N f|\|_{L^p} \leq C_p \|f\|_{L^p}$ in the case $p = 2$, and Hunt generalized the proof to cover all p with $1 < p < \infty$. This result implies that the Dirichlet sums of an L^p function do indeed converge almost everywhere when $1 < p \leq \infty$. On the other hand, this estimate fails at the endpoint $p = 1$, and there is in fact an example due to KOLMOGOROV [VI.88] of an absolutely integrable function whose Dirichlet sums are everywhere divergent. These results require a lot of harmonic analysis theory. In particular they use many decompositions of both the spatial variable and the frequency variable, keeping the Heisenberg uncertainty principle in mind. They then carefully reassemble the pieces, exploiting various manifestations of orthogonality.

To summarize, quantitative estimates such as L^p estimates on various operators provide an important route to establishing qualitative results, such as convergence of certain series or sequences. In fact there are a number of principles (notably the uniform boundedness principle and a result known as *Stein's maximal principle*) which assert that in certain circumstances this is the *only* route, in the sense that a quantitative estimate must exist in order for the qualitative result to be true.

3 Some General Themes in Harmonic Analysis: Decomposition, Oscillation, and Geometry

One feature of harmonic analysis methods is that they tend to be *local* rather than *global*. For instance, if one is analyzing a function f it is quite common to decompose it as a sum $f = f_1 + \cdots + f_k$, with each function f_i “localized” in the sense that its support (the set of values x for which $f_i(x) \neq 0$) has a small diameter. This would be called localization in the *spatial variable*. One can also localize in the *frequency variable* by applying the process to the Fourier transform \hat{f} of f . Having split f up like this, one can carry out estimates for the pieces separately and then recombine them later. One reason for this “divide and conquer” strategy is that a typical function f tends to have many different features—for example, it may be very “spiky,” “discontinuous,” or “high frequency” in some places, and “smooth” or “low frequency” in others—and it is difficult to treat all of these features at once. A well-chosen

decomposition of the function f can isolate these features from each other, so that each component has only one salient feature that could cause difficulty: the spiky part can go into one f_i , the high-frequency part into another, and so on. In reassembling the estimates from the individual components, one can use crude tools such as the triangle inequality or more refined tools, for instance those relying on some sort of orthogonality, or perhaps a clever algorithm that groups the components into manageable clusters. The main drawback of the decomposition method (other than an aesthetic one) is that it tends to give bounds that are not quite optimal; however, in many cases one is content with an estimate that differs from the best possible one by a multiplicative constant.

To give a simple example of the method of decomposition, let us consider the Fourier transform $\hat{f}(\xi)$ of a function $f : \mathbb{R} \rightarrow \mathbb{C}$, defined (for suitably nice functions f) by the formula

$$\hat{f}(\xi) = \int_{\mathbb{R}} f(x) e^{-2\pi i x \xi} dx.$$

What we can say about the size of \hat{f} , as measured by suitable norms, if we are given information about the size of f , as measured by other norms?

Here are two simple observations in response to this question. First, since the modulus of $e^{-2\pi i x \xi}$ is always equal to 1, it follows that $|\hat{f}(\xi)|$ is at most $\int_{\mathbb{R}} |f(x)| dx$. This tells us that $\|\hat{f}\|_{L^\infty} \leq \|f\|_{L^1}$, at least if $f \in L^1$. In particular, $\hat{f} \in L^\infty$. Secondly, the Plancherel theorem, a very basic fact of Fourier analysis, tells us that $\|\hat{f}\|_{L^2}$ is equal to $\|f\|_{L^2}$ if $f \in L^2$. Therefore, if f belongs to L^2 then so does \hat{f} .

We would now like to know what happens if f lies in an intermediate L^p space. In other words, what happens if $1 < p < 2$? Since L^p is not contained in either L^1 or L^2 , one cannot use either of the above two results directly. However, let us take a function $f \in L^p$ and consider what the difficulty is. The reason f may not lie in L^1 is that it may decay too slowly: for instance, the function $f(x) = (1 + |x|)^{-3/4}$ tends to zero more slowly than $1/x$ as $x \rightarrow \infty$, so its integral is infinite. However, if we raise f to the power $3/2$ we obtain the function $(1 + |x|)^{-9/8}$ which decays quickly enough to have a finite integral, so f does belong to $L^{3/2}$. Similar examples show that the reason f may fail to belong to L^2 is that it can have places where it tends to infinity slowly enough for the integral of $|f|^p$ to be finite but not slowly enough for the integral of $|f|^2$ to be finite.

Notice that these two reasons are completely different. Therefore, we can try to decompose f into two

pieces, one consisting of the part where f is large and the other consisting of the part where f is small. That is, we can choose some threshold λ and define $f_1(x)$ to be $f(x)$ when $|f(x)| < \lambda$ and 0 otherwise, and define $f_2(x)$ to be $f(x)$ when $|f(x)| \geq \lambda$ and 0 otherwise. Then $f_1 + f_2 = f$, and f_1 and f_2 are the “small part” and “large part” of f , respectively.

Because $|f_1(x)| < \lambda$ for every x , we find that

$$|f_1(x)|^2 = |f_1(x)|^{2-p} |f_1(x)|^p < \lambda^{2-p} |f_1(x)|^p.$$

Therefore, f_1 belongs to L^2 and $\|f_1\|_{L^2} \leq \lambda^{2-p} \|f_1\|_{L^p}$. Similarly, because $|f_2(x)| \geq \lambda$ whenever $f_2(x) \neq 0$, we have the inequality $|f_2(x)| \leq |f_2(x)|^p / \lambda^{p-1}$ for every x , which tells us that f_2 belongs to L^1 and that $\|f_2\|_{L^1} \leq \|f_2\|_{L^p} / \lambda^{p-1}$.

From our knowledge about the L^2 -norm of f_1 and the L^1 -norm of f_2 we can obtain upper bounds for the L^2 -norm of \hat{f}_1 and the L^∞ -norm of \hat{f}_2 , by our remarks above. By using this strategy for every λ and combining the results in a clever way, one can obtain the *Hausdorff-Young inequality*, which is the following assertion. Let p lie between 1 and 2 and let p' be the *dual exponent* of p , which is the number $p/(p-1)$. Then there is a constant C_p such that, for every function $f \in L^p$, one has the inequality $\|\hat{f}\|_{L^{p'}} \leq C_p \|f\|_{L^p}$. The particular decomposition method we have used to obtain this result is formally known as the method of *real interpolation*. It does not give the best possible value of C_p , which turns out to be $p^{1/2p} / (p')^{1/2p'}$, but that requires more delicate methods.

Another basic theme in harmonic analysis is the attempt to quantify the elusive phenomenon of *oscillation*. Intuitively, if an expression oscillates wildly, then we expect its average value to be relatively small in magnitude, since the positive and negative parts, or in the complex case the parts with a wide range of different arguments, will cancel out. For instance, if a 2π -periodic function f is smooth, then for large n the Fourier coefficient

$$\hat{f}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-inx} dx$$

will be very small since $\int_{-\pi}^{\pi} e^{-inx} dx = 0$ and the comparatively slow variation in $f(x)$ is not enough to stop the cancellation occurring. This assertion can easily be proved rigorously by repeated integration by parts. Generalizations of this phenomenon include the so-called *principle of stationary phase*, which among other things allows one to obtain precise control on the Airy function $\text{Ai}(x)$ discussed earlier. It also yields the Heisenberg uncertainty principle, which relates the

decay and smoothness of a function to the decay and smoothness of its Fourier transform.

A somewhat different manifestation of oscillation lies in the principle that if one has a sequence of functions that oscillate in different ways, then their sum should be significantly smaller than the bound that follows from the triangle inequality. Again, this is the result of cancellation that is simply not noticed by the triangle inequality. For instance, the Plancherel theorem in Fourier analysis implies, among other things, that a trigonometric polynomial $\sum_{n=-N}^N c_n e^{inx}$ has an L^2 -norm of

$$\left(\frac{1}{2\pi} \int_0^{2\pi} \left| \sum_{n=-N}^N c_n e^{inx} \right|^2 dx \right)^{1/2} = \left(\sum_{n=-N}^N |c_n|^2 \right)^{1/2}.$$

This bound (which can also be proved by direct calculation) is smaller than the upper bound of $\sum_{n=-N}^N |c_n|$ that would be obtained if we simply applied the triangle inequality to the functions $c_n e^{inx}$. This identity can be viewed as a special case of the Pythagorean theorem, together with the observation that the harmonics e^{inx} are all *orthogonal* to each other with respect to the INNER PRODUCT [III.37]

$$\langle f, g \rangle = \frac{1}{2\pi} \int_0^{2\pi} f(x) \overline{g(x)} dx.$$

This concept of orthogonality has been generalized in a number of ways. For instance, there is a more general and robust concept of “almost orthogonality,” which roughly speaking means that the inner products of a collection of functions are small but not necessarily 0.

Many arguments in harmonic analysis will, at some point, involve a combinatorial statement about certain types of geometric objects such as cubes, balls, or boxes. For instance, one useful such statement is the *Vitali covering lemma*, which asserts that, given any collection B_1, \dots, B_k of balls in Euclidean space \mathbb{R}^n , there will be a subcollection B_{i_1}, \dots, B_{i_m} of balls that are disjoint, but that nevertheless contain a significant fraction of the volume covered by the original balls. To be precise, one can choose the disjoint balls so that

$$\text{vol} \left(\bigcup_{j=1}^m B_{i_j} \right) \geq 5^{-n} \text{vol} \left(\bigcup_{j=1}^k B_j \right).$$

(The constant 5^{-n} can be improved, but this will not concern us here.) This result is obtained by a “greedy algorithm”: one picks balls one by one, at each stage choosing the largest ball among the B_j that is disjoint from all the balls already selected.

One consequence of the Vitali covering lemma is the *Hardy-Littlewood maximal inequality*, which we will

briefly describe. Given any function $f \in L^1(\mathbb{R}^n)$, any $x \in \mathbb{R}^n$, and any $r > 0$, we can calculate the average of $|f|$ in the n -dimensional sphere $B(x, r)$ of center x and radius r . Next, we can define the *maximal function* F of f by letting $F(x)$ be the largest of all these averages as r ranges over all positive real numbers. (More precisely, one takes the supremum.) Then, for each positive real number λ one can define a set X_λ to be the set of all x such that $F(x) > \lambda$. The Hardy-Littlewood maximal inequality asserts that the volume of X_λ is at most $5^n \|f\|_{L^1} / \lambda$.²

To prove it, one observes that X_λ can be covered by balls $B(x, r)$ on each of which the integral of $|f|$ is at least $\lambda \text{vol}(B(x, r))$. To this collection of balls one can then apply the Vitali covering lemma, and the result follows. The Hardy-Littlewood maximal inequality is a quantitative result, but it has as a qualitative consequence the *Lebesgue differentiation theorem*, which asserts the following. If f is any absolutely integrable function defined on \mathbb{R}^n , then for almost every $x \in \mathbb{R}^n$ the averages

$$\frac{1}{\text{vol}(B(x, r))} \int_{B(x, r)} f(y) dy$$

of f over the Euclidean balls about x tend to $f(x)$ as $r \rightarrow 0$. This example demonstrates the importance of the underlying geometry (in this case, the combinatorics of metric balls) in harmonic analysis.

Further Reading

- Stein, E. M. 1970. *Singular Integrals and Differentiability Properties of Functions*. Princeton, NJ: Princeton University Press.
- . 1993. *Harmonic Analysis*. Princeton, NJ: Princeton University Press.
- Wolff, T. H. 2003. *Lectures on Harmonic Analysis* (ed. I. Łaba and C. Shubin). University Lecture Series, volume 29. Providence, RI: American Mathematical Society.

PUP: further reading added after proofreader received proof. Please check.

IV.12 Partial Differential Equations

Sergiu Klainerman

Introduction

Partial differential equations (or PDEs) are an important class of *functional equations*: they are equations, or systems of equations, in which the unknowns are

2. This version of the Hardy-Littlewood inequality looks somewhat different from the one mentioned briefly in the previous section, but one can deduce that inequality from this one by the real interpolation method discussed earlier.

PUP: it's the balls that are disjoint, not the collection, so sentence OK?

functions of more than one variable. As a very crude analogy, PDEs are to functions as polynomial equations (such as $x^2 + y^2 = 1$, for example) are to numbers. The distinguishing feature of PDEs, as opposed to more general functional equations, is that they involve not only unknown functions, but also various *partial derivatives* of those functions, in algebraic combination with each other and with other, fixed, functions. Other important kinds of functional equations are *integral equations*, which involve various integrals of the unknown functions, and *ordinary differential equations* (ODEs), in which the unknown functions depend on only one independent variable (such as a time variable t) and the equation involves only ordinary derivatives $d/dt, d^2/dt^2, d^3/dt^3, \dots$ of these functions.

Given the immense scope of the subject the best I can hope to do is to give a very crude perspective on some of the main issues and an even cruder idea of the multitude of current research directions. The difficulty one faces in trying to describe the subject of PDEs starts with its very definition. Is it a unified area of mathematics, devoted to the study of a clearly defined set of objects (in the way that algebraic geometry studies solutions of polynomial equations or topology studies manifolds, for example), or is it rather a collection of separate fields, such as general relativity, several complex variables, or hydrodynamics, each one vast in its own right and centered on a particular, very difficult, equation or class of equations? I will attempt to argue below that, even though there are fundamental difficulties in formulating a general theory of PDEs, one can nevertheless find a remarkable unity between various branches of mathematics and physics that are centered on individual PDEs or classes of PDEs. In particular, certain ideas and methods in PDEs have turned out to be extraordinarily effective across the boundaries of these separate fields. It is thus no surprise that the most successful book ever written about PDEs did not mention PDEs in its title: it was *Methods of Mathematical Physics* by COURANT [VI.83] and HILBERT [VI.63].

As it is impossible to do full justice to such a huge subject in such limited space I have been forced to leave out many topics and relevant details; in particular, I have said very little about the fundamental issue of breakdown of solutions, and there is no discussion of the main open problems in PDEs. A longer and more detailed version of the article, which includes these topics, can be found at

<http://press.princeton.edu/???>

1 Basic Definitions and Examples

The simplest example of a PDE is the LAPLACE EQUATION [I.3 §5.4]

$$\Delta u = 0. \quad (1)$$

Here, Δ is the *Laplacian*, that is, the *differential operator* that transforms functions $u = u(x_1, x_2, x_3)$ defined from \mathbb{R}^3 to \mathbb{R} according to the rule

$$\begin{aligned} \Delta u(x_1, x_2, x_3) \\ = \partial_1^2 u(x_1, x_2, x_3) + \partial_2^2 u(x_1, x_2, x_3) + \partial_3^2 u(x_1, x_2, x_3), \end{aligned}$$

where $\partial_1, \partial_2, \partial_3$ are standard shorthand for the partial derivatives $\partial/\partial x_1, \partial/\partial x_2, \partial/\partial x_3$. (We will use this shorthand throughout the article.) Two other fundamental examples (also described in [I.3 §5.4]) are the *heat equation* and the *wave equation*:

$$-\partial_t u + k\Delta u = 0, \quad (2)$$

$$-\partial_t^2 u + c^2 \Delta u = 0. \quad (3)$$

In each case one is asked to find a function u that satisfies the corresponding equations. For the Laplace equation u will depend on x_1, x_2 , and x_3 , and for the other two it will depend on t as well. Observe that equations (2) and (3) again involve the symbol Δ , but also partial derivatives with respect to the time variable t . The constants k (which is positive) and c are fixed and represent the rate of diffusion and the speed of light, respectively. However, from a mathematical point of view they are not important, since if $u(t, x_1, x_2, x_3)$ is a solution of (3), for example, then $v(t, x_1, x_2, x_3) = u(t, x_1/c, x_2/c, x_3/c)$ satisfies the same equation with $c = 1$. Thus, when one is studying the equations one can set these constants to be 1. Both equations are called *evolution equations* because they are supposed to describe the change of a particular physical object as the time parameter t varies. Observe that (1) can be interpreted as a particular case of both (2) and (3): if $u = u(t, x_1, x_2, x_3)$ is a solution of either (2) or (3) that is independent of t , then $\partial_t u = 0$, so u must satisfy (1).

In all three examples mentioned above, we tacitly assume that the solutions we are looking for are sufficiently differentiable for the equations to make sense. As we shall see later, one of the important developments in the theory of PDEs was the study of more refined notions of solutions, such as DISTRIBUTIONS [III.18], which require only *weak* versions of differentiability.

Here are some further examples of important PDEs. The first is THE SCHRÖDINGER EQUATION [III.85],

$$i\partial_t u + k\Delta u = 0, \quad (4)$$

PUP: 'but also' has to stay as a contrast is being expressed with the Laplace equation. OK?

where u is a function from $\mathbb{R} \times \mathbb{R}^3$ to \mathbb{C} . This equation describes the quantum evolution of a massive particle, $k = \hbar/2m$, where $\hbar > 0$ is Planck's constant and m is the mass of the particle. As with the heat equation, one can set k to equal 1 after a simple change of variables. Though the equation is formally very similar to the heat equation, it has very different qualitative behavior. This illustrates an important general point about PDEs: that small changes in the form of an equation can lead to very different properties of solutions.

A further example is the Klein-Gordon equation

$$-\partial_t^2 u + c^2 \Delta u - \left(\frac{mc^2}{\hbar}\right)^2 u = 0. \quad (5)$$

This is the relativistic counterpart to the Schrödinger equation: the parameter m has the physical interpretation of mass and mc^2 has the physical interpretation of rest energy (reflecting Einstein's famous equation $E = mc^2$). One can normalize the constants c and mc^2/\hbar so that they both equal 1 by applying a suitable change of variables to time and space.

Though all five equations mentioned above first appeared in connection with specific physical phenomena, such as heat transfer for (2) and propagation of electromagnetic waves for (3), they have, miraculously, a range of relevance far beyond their original applications. In particular there is no reason to restrict their study to three space dimensions: it is very easy to generalize them to similar equations in n variables x_1, x_2, \dots, x_n .

All the PDEs listed so far obey a simple but fundamental property called the *principle of superposition*: if u_1 and u_2 are two solutions to one of these equations, then any linear combination $a_1 u_1 + a_2 u_2$ of these solutions is also a solution. In other words, the space of all solutions is a VECTOR SPACE [I.3 §2.3]. Equations that obey this property are known as *homogeneous linear equations*. If the space of solutions is an affine space (that is, a translate of a vector space) rather than a vector space, we say that the PDE is an *inhomogeneous linear equation*; a good example is *Poisson's equation*:

$$\Delta u = f, \quad (6)$$

where $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ is a function that is given to us and $u : \mathbb{R}^3 \rightarrow \mathbb{R}$ is the unknown function. Equations that are neither homogeneous linear nor inhomogeneous linear are known as *nonlinear*. The following equation, the MINIMAL-SURFACE EQUATION [III.96 §3.1], is manifestly

nonlinear:

$$\partial_1 \left(\frac{\partial_1 u}{(1 + |\partial_1 u|^2 + |\partial_2 u|^2)^{1/2}} \right) + \partial_2 \left(\frac{\partial_2 u}{(1 + |\partial_1 u|^2 + |\partial_2 u|^2)^{1/2}} \right) = 0. \quad (7)$$

The graphs of solutions $u : \mathbb{R}^2 \rightarrow \mathbb{R}$ of this equation are area-minimizing surfaces (like soap films).

Equations (1), (2), (3), (4), (5) are not just linear: they are all examples of *constant-coefficient linear equations*. This means that they can be expressed in the form

$$\mathcal{P}[u] = 0, \quad (8)$$

where \mathcal{P} is a differential operator that involves linear combinations, with constant real or complex coefficients, of mixed partial derivatives of f . (Such operators are called *constant-coefficient linear differential operators*.) For instance, in the case of the Laplace equation (1), \mathcal{P} is simply the Laplacian Δ , while for the wave equation (3), \mathcal{P} is the *d'Alembertian*

$$\mathcal{P} = \square = -\partial_t^2 + \partial_1^2 + \partial_2^2 + \partial_3^2.$$

The characteristic feature of linear constant-coefficient operators is *translation invariance*. Roughly speaking, this means that if you translate a function u , then you translate $\mathcal{P}u$ in the same way. More precisely, if $v(x)$ is defined to be $u(x - a)$ (so the value of u at x becomes the value of v at $x + a$; note that x and a belong to \mathbb{R}^3 here), then $\mathcal{P}v(x)$ is equal to $\mathcal{P}u(x - a)$. As a consequence of this basic fact we infer that solutions to the homogeneous, linear, constant-coefficient equation (8) are still solutions when translated.

Since symmetries play such a fundamental role in PDEs we should stop for a moment to make a general definition. A symmetry of a PDE is any invertible operation $T : u \mapsto T(u)$ from functions to functions that preserves the space of solutions, in the sense that u solves the PDE if and only if $T(u)$ solves the same PDE. A PDE with this property is then said to be *invariant* under the symmetry T . The symmetry T is often a linear operation, though this does not have to be the case. The composition of two symmetries is again a symmetry, as is the inverse of a symmetry, and so it is natural to view a collection of symmetries as forming a GROUP [I.3 §2.1] (which is typically a finite- or infinite-dimensional LIE GROUP [III.50 §1]).

Because the translation group is intimately connected with THE FOURIER TRANSFORM [III.27] (indeed, the latter can be viewed as the representation theory of the former), this symmetry strongly suggests that

Fourier analysis should be a useful tool to solve constant-coefficient PDEs, and this is indeed the case.

Our basic constant-coefficient linear operators, the Laplacian Δ and the d'Alembertian \square , are formally similar in many respects. The Laplacian is fundamentally associated with the geometry of EUCLIDEAN SPACE [I.3 §6.2] \mathbb{R}^3 and the d'Alembertian is similarly associated with the geometry of MINKOWSKI SPACE [I.3 §6.8] \mathbb{R}^{1+3} . This means that the Laplacian commutes with all the rigid motions of the Euclidean space \mathbb{R}^3 , while the d'Alembertian commutes with the corresponding class of *Poincaré transformations* of Minkowski space-time. In the former case this simply means that invariance applies to all transformations of \mathbb{R}^3 that preserve the Euclidean distances between points. In the case of the wave equation, the Euclidean distance has to be replaced by the *spacetime distance* between points (which would be called *events* in the language of relativity): if $P = (t, x_1, x_2, x_3)$ and $Q = (s, y_1, y_2, y_3)$, then the distance between them is given by the formula

$$d_M(P, Q)^2 = -(t - s)^2 + (x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2.$$

As a consequence of this basic fact we infer that all solutions to the wave equation (3) are invariant under translations and LORENTZ TRANSFORMATIONS [I.3 §6.8].

Our other evolution equations (2) and (4) are clearly invariant under rotations of the space variables $x = (x^1, x^2, x^3) \in \mathbb{R}^3$, when t is fixed. They are also *Galilean invariant*, which means, in the particular case of the Schrödinger equation (4), that whenever $u = u(t, x)$ is a solution so is the function $u_v(t, x) = e^{i(x \cdot v)} e^{it|v|^2} (t, x - vt)$ for any vector $v \in \mathbb{R}^3$.

Poisson's equation (6), on the other hand, is an example of a *constant-coefficient inhomogeneous linear equation*, which means that it takes the form

$$\mathcal{P}[u] = f \quad (9)$$

for some constant-coefficient linear differential operator \mathcal{P} and known function f . To solve such an equation requires one to understand the invertibility or otherwise of the linear operator \mathcal{P} : if it is invertible then u will equal $\mathcal{P}^{-1}f$, and if it is not invertible then either there will be no solution or there will be infinitely many solutions. Inhomogeneous equations are closely related to their homogeneous counterpart; for instance, if u_1, u_2 both solve the inhomogeneous equation (9) with the same inhomogeneous term f , then their difference $u_1 - u_2$ solves the corresponding homogeneous equation (8).

Linear homogeneous PDEs satisfy the principle of superposition but they do not have to be translation invariant. For example, suppose that we modify the heat equation (2) so that the coefficient k is no longer constant but rather an arbitrary, positive, smooth function of (x_1, x_2, x_3) . Such an equation models the flow of heat in a medium in which the rate of diffusion varies from point to point. The corresponding space of solutions is not translation invariant (which is not surprising as the medium in which the heat flows is not translation invariant). Equations like this are called *linear equations with variable coefficients*. It is more difficult to solve them and describe their qualitative features than it is for constant-coefficient equations. (See, for example, STOCHASTIC PROCESSES [IV.24 §5.2] for an approach to equations of type (2) with variable k .) Finally, nonlinear equations such as (7) can often still be written in the form (8), but the operator \mathcal{P} is now a *nonlinear* differential operator. For instance, the relevant operator for (7) is given by the formula

$$\mathcal{P}[u] = \sum_{i=1}^2 \partial_i \left(\frac{1}{(1 + |\partial u|^2)^{1/2}} \partial_i u \right),$$

where $|\partial u|^2 = (\partial_1 u)^2 + (\partial_2 u)^2$. Operators such as these are clearly not linear. However, because they are ultimately constructed from algebraic operations and partial derivatives, both of which are “local” operations, we observe the important fact that \mathcal{P} is at least still a “local” operator. More precisely, if u_1 and u_2 are two functions that agree on some open set D , then the expressions $\mathcal{P}[u_1]$ and $\mathcal{P}[u_2]$ also agree on this set. In particular, if $\mathcal{P}[0] = 0$ (as is the case in our example), then whenever u vanishes on a domain, $\mathcal{P}[u]$ will also vanish on that domain.

So far we have tacitly assumed that our equations take place in the whole of a space such as \mathbb{R}^3 , $\mathbb{R}^+ \times \mathbb{R}^3$, or $\mathbb{R} \times \mathbb{R}^3$. In reality one is often restricted to a fixed domain of that space. Thus, for example, equation (1) is usually studied on a bounded open domain of \mathbb{R}^3 subject to a specified *boundary condition*. Here are some basic examples of boundary conditions.

Example. The *Dirichlet problem* for Laplace's equation on an open domain of $D \subset \mathbb{R}^3$ is the problem of finding a function u that behaves in a prescribed way on the boundary of D and obeys the Laplace equation inside.

More precisely, one specifies a continuous function $u_0 : \partial D \rightarrow \mathbb{R}$ and looks for a continuous function u , defined on the closure \bar{D} of D , that is twice continu-

ously differentiable inside D and solves the equations

$$\left. \begin{aligned} \Delta u(x) &= 0 && \text{for all } x \in D, \\ u(x) &= u_0(x) && \text{for all } x \in \partial D. \end{aligned} \right\} \quad (10)$$

A basic result in PDEs asserts that if the domain D has a sufficiently smooth boundary, then there is exactly one solution to the problem (10) for any prescribed function u_0 on the boundary ∂D .

Example. The *Plateau problem* is the problem of finding the surface of minimal total area that bounds a given curve.

When the surface is the graph of a function u on some suitably smooth domain D , in other words a set of the form $\{(x, y, u(x, y)) : (x, y) \in D\}$, and the bounding curve is the graph of a function u_0 over the boundary ∂D of D , then this problem turns out to be equivalent to the Dirichlet problem (10), but with the linear equation (1) replaced by the nonlinear equation (7). For the above equations, it is also often natural to replace the Dirichlet boundary condition $u(x) = u_0(x)$ on the boundary ∂D with another boundary condition, such as the *Neumann boundary condition* $n(x) \cdot \nabla_x u(x) = u_1(x)$ on ∂D , where $n(x)$ is the outward normal (of unit length) to D at x . Generally speaking, Dirichlet boundary conditions correspond to “absorbing” or “fixed” barriers in physics, whereas Neumann boundary conditions correspond to “reflecting” or “free” barriers.

Natural boundary conditions can also be imposed for our evolution equations (2)–(4). The simplest one is to prescribe the values of u when $t = 0$. We can think of this more geometrically. We are prescribing the values of u at each spacetime point of form $(0, x, y, z)$, and the set of all such points is a hyperplane in \mathbb{R}^{1+3} : it is an example of an *initial time surface*.

Example. The *Cauchy problem* (or *initial-value problem*, sometimes abbreviated to IVP) for the heat equation (2) asks for a solution $u : \mathbb{R}^+ \times \mathbb{R}^3 \rightarrow \mathbb{R}$ on the spacetime domain $\mathbb{R}^+ \times \mathbb{R}^3 = \{(t, x) : t > 0, x \in \mathbb{R}^3\}$, which equals a prescribed function $u_0 : \mathbb{R}^3 \rightarrow \mathbb{R}$ on the initial time surface $\{0\} \times \mathbb{R}^3 = \partial(\mathbb{R}^+ \times \mathbb{R}^3)$.

In other words, the Cauchy problem asks for a sufficiently smooth function u , defined on the closure of $\mathbb{R}^+ \times \mathbb{R}^3$ and taking values in \mathbb{R} , that satisfies the conditions

$$\left. \begin{aligned} -\partial_t u(t, x) + k\Delta u(t, x) &= 0 \\ &\text{for every } (t, x) \in \mathbb{R}^+ \times \mathbb{R}^3, \\ u(0, x) &= u_0(x) \text{ for every } x \in \mathbb{R}^3. \end{aligned} \right\} \quad (11)$$

The function u_0 is often referred to as the *initial conditions*, or *initial data*, or just *data*, for the problem. Under suitable smoothness and decay conditions, one can show that this equation has exactly one solution u for each choice of data u_0 . Interestingly, this assertion fails if one replaces the *future* domain $\mathbb{R}^+ \times \mathbb{R}^3 = \{(t, x) : t > 0, x \in \mathbb{R}^3\}$ by the *past* domain $\mathbb{R}^- \times \mathbb{R}^3 = \{(t, x) : t < 0, x \in \mathbb{R}^3\}$.

A similar formulation of the IVP holds for the Schrödinger equation (4), though in this case we can solve both to the past and to the future. However, in the case of the wave equation (3) we need to specify not just the initial *position* $u(0, x) = u_0(x)$ on the initial time surface $t = 0$, but also an initial *velocity* $\partial_t u(0, x) = u_1(x)$, since equation (3) (unlike (2) or (4)) cannot formally determine $\partial_t u$ in terms of u . One can construct unique smooth solutions (both to the future and to the past of the initial hyperplane $t = 0$) to the IVP for (3) for very general smooth initial conditions u_0, u_1 .

Many other boundary-value problems are possible. For instance, when analyzing the evolution of a wave in a bounded domain D (such as a sound wave), it is natural to work with the spacetime domain $\mathbb{R} \times D$ and prescribe *both* Cauchy data (on the initial boundary $0 \times D$) and Dirichlet or Neumann data (on the spatial boundary $\mathbb{R} \times \partial D$). On the other hand, when the physical problem under consideration is the evolution of a wave outside a bounded obstacle (for example, an electromagnetic wave), one considers instead the evolution in $\mathbb{R} \times (\mathbb{R}^3 \setminus D)$ with a boundary condition on D .

The choice of boundary condition and initial conditions for a given PDE is very important. For equations of physical interest these arise naturally from the context in which they are derived. For example, in the case of a vibrating string, which is described by solutions of the one-dimensional wave equation $\partial_t^2 u - \partial_x^2 u = 0$ in the domain $(a, b) \times \mathbb{R}$, the initial conditions $u = u_0$ and $\partial_t u = u_1$ at $t = t_0$ amount to specifying the original position and velocity of the string. The boundary condition $u(a) = u(b) = 0$ is what tells us that the two ends of the string are fixed.

So far we have considered just *scalar* equations. These are equations where there is only one unknown function u , which takes values either in the real numbers \mathbb{R} or in the complex numbers \mathbb{C} . However, many important PDEs involve either multiple unknown scalar functions or (equivalently) functions that take values in a multidimensional vector space such as \mathbb{R}^m . In such cases, we say that we have a *system* of PDEs. An

important example of a system is that of the CAUCHY-RIEMANN EQUATIONS [I.3 §5.6]:

$$\partial_1 u_2 - \partial_2 u_1 = 0, \quad \partial_1 u_1 + \partial_2 u_2 = 0, \quad (12)$$

where $u_1, u_2 : \mathbb{R}^2 \rightarrow \mathbb{R}$ are real-valued functions on the plane. It was observed by CAUCHY [VI.29] that a complex function $w(x + iy) = u_1(x, y) + iu_2(x, y)$ is HOLOMORPHIC [I.3 §5.6] if and only if its real and imaginary parts u_1, u_2 satisfy the system (12). This system can still be represented in the form of a constant-coefficient linear PDE (8), but u is now a vector $\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$, and \mathcal{P} is not a scalar differential operator, but rather a *matrix* of operators $\begin{pmatrix} -\partial_2 & \partial_1 \\ \partial_1 & \partial_2 \end{pmatrix}$.

The system (12) contains two equations and two unknowns. This is the standard situation for a *determined system*. Roughly speaking, a system is called *overdetermined* if it contains more equations than unknowns and *underdetermined* if it contains fewer equations than unknowns. Underdetermined equations typically have infinitely many solutions for any given set of prescribed data; conversely, overdetermined equations tend to have no solutions at all, unless some additional *compatibility conditions* are imposed on the prescribed data.

Observe also that the Cauchy-Riemann operator \mathcal{P} has the following remarkable property:

$$\mathcal{P}^2[u] = \mathcal{P}[\mathcal{P}[u]] = \begin{pmatrix} \Delta u_1 \\ \Delta u_2 \end{pmatrix}.$$

Thus \mathcal{P} can be viewed as a square root of the two-dimensional Laplacian Δ . One can define a similar type of square root for the Laplacian in higher dimensions and, more surprisingly, even for the d'Alembertian operator \square in \mathbb{R}^{1+3} . To achieve this we need to have four 4×4 complex matrices $\gamma^0, \gamma^1, \gamma^3, \gamma^4$ that satisfy the property

$$\gamma^\alpha \gamma^\beta + \gamma^\beta \gamma^\alpha = -2m^{\alpha\beta} I.$$

Here, I is the unit 4×4 matrix and $m^{\alpha\beta} = \frac{1}{2}$ when $\alpha = \beta = 1, -\frac{1}{2}$ when $\alpha = \beta \neq 1$, and 0 otherwise. Using the γ matrices we can introduce the *Dirac operator* as follows. If $u = (u_1, u_2, u_3, u_4)$ is a function in \mathbb{R}^{1+3} with values in \mathbb{C}^4 , then we set $\mathcal{D}u = i\gamma^\alpha \partial_\alpha u$. It is easy to check that, indeed, $\mathcal{D}^2 u = \square u$. The equation

$$\mathcal{D}u = ku \quad (13)$$

is called the *Dirac equation* and it is associated with a free, massive, relativistic particle such as an electron.

One can extend the concept of a PDE further to cover unknowns that are not, strictly speaking, functions

taking values in a vector space, but are instead sections of a VECTOR BUNDLE [IV.6 §5], or perhaps a map from one MANIFOLD [I.3 §6.9] to another; such generalized PDEs play an important role in geometry and modern physics. A fundamental example is given by the EINSTEIN FIELD EQUATIONS [IV.13]. In the simplest, “vacuum,” case, they take the form

$$\text{Ric}(g) = 0, \quad (14)$$

where $\text{Ric}(g)$ is the RICCI CURVATURE [III.80] tensor of the spacetime manifold $M = (M, g)$. In this case the spacetime metric itself is the unknown to be solved for. One can often reduce such equations *locally* to more traditional PDE systems by selecting a suitable choice of coordinates, but the task of selecting a “good” choice of coordinates, and working out how different choices are compatible with each other, is a nontrivial and important one. Indeed, the task of selecting a good set of coordinates in order to solve a PDE can end up being a significant PDE problem in its own right.

PDEs are ubiquitous throughout mathematics and science. They provide the basic mathematical framework for some of the most important physical theories: elasticity, hydrodynamics, electromagnetism, general relativity, and nonrelativistic quantum mechanics, for example. The more modern relativistic quantum field theories lead, in principle, to equations in an infinite number of unknowns, which lie beyond the scope of PDEs. Yet, even in that case, the basic equations preserve the locality property of PDEs. Moreover, the starting point of a QUANTUM FIELD THEORY [IV.17 §2.1.4] is always a classical field theory, which is described by systems of PDEs. This is the case, for example, in the standard model of weak and strong interactions, which is based on the so-called Yang-Mills-Higgs field theory. If we also include the ordinary differential equations of classical mechanics, which can be viewed as one-dimensional PDEs, we see that essentially all of physics is described by differential equations. As examples of PDEs underlying some of our most basic physical theories we refer to the articles that discuss THE EULER AND NAVIER-STOKES EQUATIONS [III.23], THE HEAT EQUATION [III.36], THE SCHRÖDINGER EQUATION [III.85], and THE EINSTEIN EQUATIONS [IV.13].

An important feature of the main PDEs is their apparent universality. Thus, for example, the wave equation, first introduced by D'ALEMBERT [VI.20] to describe the motion of a vibrating string, was later found to be connected with the propagation of sound and electromagnetic waves. The heat equation, first introduced by

FOURIER [VI.25] to describe heat propagation, appears in many other situations in which dissipative effects play an important role. The same can be said about the Laplace equation, the Schrödinger equation, and many other basic equations.

It is even more surprising that equations that were originally introduced to describe specific physical phenomena have played a fundamental role in several areas of mathematics that are considered to be “pure,” such as complex analysis, differential geometry, topology, and algebraic geometry. Complex analysis, for example, which studies the properties of holomorphic functions, can be regarded as the study of solutions to the Cauchy–Riemann equations (12) in a domain of \mathbb{R}^2 . Hodge theory is based on studying the space of solutions to a class of linear systems of PDEs on manifolds that generalize the Cauchy–Riemann equations: it plays a fundamental role in topology and algebraic geometry. THE ATIYAH–SINGER INDEX THEOREM [V.2] is formulated in terms of a special class of linear PDEs on manifolds, related to the Euclidean version of the Dirac operator. Important geometric problems can be reduced to finding solutions to specific PDEs, typically nonlinear. We have already seen one example: the Plateau problem of finding surfaces of minimal total area that pass through a given curve. Another striking example is the UNIFORMIZATION THEOREM [V.37] in the theory of surfaces, which takes a compact Riemannian surface S (a two-dimensional surface with a RIEMANNIAN METRIC [I.3 §6.10]) and, by solving the PDE

$$\Delta_S u + e^{2u} = K \quad (15)$$

(which is a nonlinear variant of the Laplace equation (1)), *uniformizes* the metric so that it is “equally curved” at all points on the surface (or, more precisely, has constant SCALAR CURVATURE [III.80]) without changing the *conformal class* of the metric (i.e., without distorting any of the angles subtended by curves on the surface). This theorem is of fundamental importance to the theory of such surfaces: in particular, it allows one to give a topological classification of compact surfaces in terms of a single number $\chi(S)$, which is called the EULER CHARACTERISTIC [I.4 §2.2] of the surface S . The three-dimensional analogue of the uniformization theorem, the GEOMETRIZATION CONJECTURE [IV.7 §2.4] of Thurston, has recently been established by Perelman, who did so by solving yet another PDE; in this case, the equation is the RICCI FLOW [III.80] equation

$$\partial_t g = 2 \operatorname{Ric}(g), \quad (16)$$

which can be transformed into a nonlinear version of the heat equation (2) after a carefully chosen change of coordinates. The proof of the geometrization conjecture is a decisive step toward the total classification of all three-dimensional compact manifolds, in particular establishing the well-known POINCARÉ CONJECTURE [IV.7 §2.4]. To overcome the many technical details in establishing this conjecture, one needs to make a detailed qualitative analysis of the behavior of solutions to the Ricci flow equation, a task which requires just about all the advances made in geometric PDEs in the last hundred years.

Finally, we note that PDEs arise not only in physics and geometry but also in many fields of applied science. In engineering, for example, one often wants to *control* some feature of the solution u to a PDE by carefully selecting whatever components of the given data one can directly influence; consider, for instance, how a violinist controls the solution to the vibrating string equation (closely related to (3)) by modulating the force and motion of a bow on that string in order to produce a beautiful sound. The mathematical theory dealing with these types of issues is called *control theory*.

When dealing with complex physical systems, one cannot possibly have complete information about the state of the system at any given time. Instead, one often makes certain randomness assumptions about various factors that influence it. This leads to the very important class of equations called *stochastic differential equations* (SDEs), where one or more components of the equation involve a RANDOM VARIABLE [III.73 §4] of some sort. An example of this is in the BLACK–SCHOLES MODEL [VII.9 §2] in mathematical finance. A general discussion of SDEs can be found in STOCHASTIC PROCESSES [IV.24 §6].

The plan for the rest of this article is as follows. In section 2 I shall describe some of the basic notions and achievements of the general theory of PDEs. The main point I want to make here is that, in contrast with ordinary differential equations, for which a general theory is both possible and useful, partial differential equations do not lend themselves to a useful general theoretical treatment because of some important obstructions that I shall try to describe. One is thus forced to discuss special classes of equations such as *elliptic*, *parabolic*, *hyperbolic*, and *dispersive equations*. In section 3 I will try to argue that, despite the impossibility of developing a useful general theory that encompasses all, or most, of the important examples, there is nevertheless an impressive unifying body of concepts

and methods for dealing with various basic equations, and this gives PDEs the feel of a well-defined area of mathematics. In section 4 I develop this further by trying to identify some common features in the derivation of the main equations that are dealt with in the subject. An additional source of unity for PDEs is the central role played by the issues of *regularity* and *breakdown* of solutions, which is discussed only briefly here. In the final section we shall discuss some of the main goals that can be identified as driving the subject.

2 General Equations

One might expect, after looking at other areas of mathematics such as algebraic geometry or topology, that there was a very general theory of PDEs that could be specialized to various specific cases. As I shall argue below, this point of view is seriously flawed and very much out of fashion. It does, however, have important merits, which I hope to illustrate in this section. I shall avoid giving formal definitions and focus instead on representative examples. The reader who wants more precise definitions can consult the online version of this article.

For simplicity we shall look mostly at *determined* systems of PDEs. The simplest distinction, which we have already made, is between scalar equations, such as (1)–(5), which consist of only one equation and one unknown, and systems of equations, such as (12) and (13). Another simple but important concept is that of the *order* of a PDE, which is defined to be the highest derivative that appears in the equation; this concept is analogous to that of the *degree* of a polynomial. For instance, the five basic equations (1)–(5) listed earlier are second order in space, although some (such as (2) or (4)) are only first order in time. Equations (12) and (13), as well as the Maxwell equations, are first order.¹

We have seen that PDEs can be divided into linear and nonlinear equations, with the linear equations being divided further into constant-coefficient and variable-coefficient equations. One can also divide nonlinear PDEs into several further classes depending on the “strength” of the nonlinearity. At one end of the scale, a *semilinear* equation is one in which all the nonlinear components of the equation have strictly lower order than the linear components. For instance, equation (15) is semilinear, because the nonlinear component e^u is

of zero order, i.e., it contains no derivatives, whereas the linear component $\Delta_S u$ is of second order. These equations are close enough to being linear that they can often be effectively viewed as perturbations of a linear equation. A more strongly nonlinear class of equations is that of *quasilinear equations*, in which the highest-order derivatives of u appear in the equation only in a linear manner but the coefficients attached to those derivatives may depend in some nonlinear manner on lower-order derivatives. For instance, the second-order equation (7) is quasilinear, because if one uses the product rule to expand the equation, then it takes the quasilinear form

$$F_{11}(\partial_1 u, \partial_2 u) \partial_1^2 u + F_{12}(\partial_1 u, \partial_2 u) \partial_1 \partial_2 u + F_{22}(\partial_1 u, \partial_2 u) \partial_2^2 u = 0$$

for some explicit algebraic functions F_{11}, F_{12}, F_{22} of the lower-order derivatives of u . While quasilinear equations can still sometimes be analyzed by perturbative techniques, this is generally more difficult to accomplish than it is for an analogous semilinear equation. Finally, we have *fully nonlinear equations*, which exhibit no linearity properties whatsoever. A typical example is the *Monge-Ampere equation*

$$\det(D^2 u) = F(x, u, Du),$$

where $u : \mathbb{R}^n \rightarrow \mathbb{R}$ is the unknown function, Du is the GRADIENT [I.3 §5.3] of u , $D^2 u = (\partial_i \partial_j u)_{1 \leq i, j \leq n}$ is the *Hessian matrix* of u , and $F : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}$ is a given function. This equation arises in many geometric contexts, ranging from manifold-embedding problems to the complex geometry of CALABI-YAU MANIFOLDS [III.6]. Fully nonlinear equations are among the most difficult and least well-understood of all PDEs.

Remark. Most of the basic equations of physics, such as the Einstein equations, are quasilinear. However, fully nonlinear equations arise in the theory of characteristics of linear PDEs, which we discuss below, and also in geometry.

2.1 First-Order Scalar Equations

It turns out that first-order scalar PDEs in any number of dimensions can be reduced to systems of first-order ODEs. As a simple illustration of this important fact consider the following equation in two space dimensions:

$$a^1(x^1, x^2) \partial_1 u(x^1, x^2) + a^2(x^1, x^2) \partial_2 u(x^1, x^2) = f(x^1, x^2), \quad (17)$$

1. There is a simple trick, well-known in ordinary differential equations, for converting higher-order equations into a lower-order (or even first-order) system of equations by increasing the number of unknowns. See the discussion in DYNAMICS [IV.14 §1.2].

PUP: another good spot by the proofreader here - thanks!

where a^1, a^2, f are given real functions in the variables $x = (x^1, x^2) \in \mathbb{R}^2$. We associate with (17) the first-order 2×2 system

$$\left. \begin{aligned} \frac{dx^1}{ds}(s) &= a^1(x^1(s), x^2(s)), \\ \frac{dx^2}{ds}(s) &= a^2(x^1(s), x^2(s)). \end{aligned} \right\} \quad (18)$$

To simplify matters, let us assume that $f = 0$.

Suppose now that $x(s) = (x^1(s), x^2(s))$ is a solution of (18), and let us consider how $u(x^1(s), x^2(s))$ varies as s varies. By the chain rule we know that

$$\frac{d}{ds}u = \partial_1 u \frac{dx^1}{ds} + \partial_2 u \frac{dx^2}{ds},$$

and equations (17) and (18) imply that this equals zero (by our assumption that $f = 0$). In other words, any solution $u = u(x^1, x^2)$ of (17) with $f = 0$ is constant along any parametrized curve of the form $x(s) = (x^1(s), x^2(s))$ that satisfies (18).

Thus, in principle, if we know the solutions to (18), which are called *characteristic curves* for the equation (17), then we can find all solutions to (17). I say “in principle” because, in general, the nonlinear system (18) is not so easy to solve. Nevertheless, ODEs are simpler to deal with, and the fundamental theorem of ODEs, which we will discuss later in this section, allows us to solve (18) at least locally and for a small interval in s .

The fact that u is constant along characteristic curves allows us to obtain important qualitative information even when we cannot find explicit solutions. For example, suppose that the coefficients a^1, a^2 are smooth (or real analytic) and that the initial data is smooth (or real analytic) everywhere on the set \mathcal{H} where it is defined, except at some point x_0 where it is discontinuous. Then the solution u remains smooth (or real analytic) at all points except along the characteristic curve Γ that starts at x_0 , or, in other words, along the solution to (18) that satisfies the initial condition $x(0) = x_0$. That is, the discontinuity at x_0 propagates precisely along Γ . We see here the simplest manifestation of an important principle, which we shall explain in more detail later: *singularities of solutions to PDEs propagate along characteristics* (or, more generally, hypersurfaces).

One can generalize equation (17) to allow the coefficients a_1, a_2 , and f to depend not only on $x = (x^1, x^2)$ but also on u :

$$a^1(x, u(x))\partial_1 u(x) + a^2(x, u(x))\partial_2 u(x) = f(x, u(x)). \quad (19)$$

The associated *characteristic system* becomes

$$\left. \begin{aligned} \frac{dx^1}{ds}(s) &= a^1(x(s), u(s, x(s))), \\ \frac{dx^2}{ds}(s) &= a^2(x(s), u(s, x(s))). \end{aligned} \right\} \quad (20)$$

As a special example of (19) consider the scalar equation in two space dimensions,

$$\partial_t u + u\partial_x u = 0, \quad u(0, x) = u_0(x), \quad (21)$$

which is called the *Burger equation*. Here we have set $a^1(x, u(x)) = 1$ and $a^2(x, u(x)) = u(x)$. With this choice of a^1, a^2 , we can take $x^1(s)$ to be s in (20). Then, renaming $x^2(s)$ as $x(s)$, we derive the *characteristic equation* in the form

$$\frac{dx}{ds}(s) = u(s, x(s)). \quad (22)$$

For any given solution u of (21) and any characteristic curve $(s, x(s))$ we have $(d/ds)u(s, x(s)) = 0$. Thus, in principle, knowing the solutions to (22) should allow us to determine the solutions to (21). However, this argument seems worryingly circular, since u itself appears in (22).

To see how this difficulty can be circumvented, consider the IVP for (21): that is, look for solutions that satisfy $u(0, x) = u_0(x)$. Consider an associated characteristic curve $x(s)$ such that, initially, $x(0) = x_0$. Then, since u is constant along the curve, we must have $u(s, x(s)) = u_0(x_0)$. Hence, going back to (22), we infer that $dx/ds = u_0(x_0)$ and thus $x(s) = x_0 + su_0(x_0)$. We thus deduce that

$$u(s, x_0 + su_0(x_0)) = u_0(x_0), \quad (23)$$

which implicitly gives us the form of the solution u . We see once more, from (23), that if the initial data is smooth (or real analytic) everywhere except at a point x_0 of the line $t = 0$, then the corresponding solution is also smooth (or real analytic) everywhere in a small neighborhood V of x_0 , except along the characteristic curve that begins at x_0 . The smallness of V is necessary here because new singularities can form at large scales. Indeed, u has to be constant along the lines $x + su_0(x)$, whose slopes depend on $u_0(x)$. At a point where these lines cross we would obtain different values of u , which is impossible unless u becomes singular by this point. This blow-up phenomenon occurs for any smooth, nonconstant initial data u_0 .

Remark. There is an important difference between the linear equation (17) and the quasilinear equation (19). The characteristics of the first depend only on the coefficients $a^1(x), a^2(x)$, while the characteristics of the

second depend explicitly on a particular solution u of the equation. In both cases, singularities can only propagate along the characteristic curves of the equation. For nonlinear equations, however, new singularities can form at large distance scales, whatever the smoothness of the initial data.

The above procedure extends to fully nonlinear scalar equations in \mathbb{R}^d such as the *Hamilton-Jacobi equation*

$$\partial_t u + H(x, Du) = 0, \quad u(0, x) = u_0(x), \quad (24)$$

where $u : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}$ is the unknown function, Du is the gradient of u , and the HAMILTONIAN [III.35] $H : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ and the initial data $u_0 : \mathbb{R}^d \rightarrow \mathbb{R}$ are given. For instance, the *eikonal equation* $\partial_t u = |Du|$ is a special instance of a Hamilton-Jacobi equation. We associate with (24) the ODE system

$$\left. \begin{aligned} \frac{dx^i}{dt} &= \frac{\partial}{\partial p_i} H(x(t), p(t)), \\ \frac{dp_i}{dt} &= -\frac{\partial}{\partial x^i} H(x(t), p(t)), \end{aligned} \right\} \quad (25)$$

where i runs from 1 to d . The equations (25) are known as a *Hamiltonian system* of ODEs. The relationship between this system and the corresponding Hamilton-Jacobi equation is a little more involved than in the cases discussed above. Briefly, we can construct a solution u to (24) based only on the knowledge of the solutions $(x(t), p(t))$ to (25), which are called the *bicharacteristic curves* of the nonlinear PDE. Once again, singularities can only propagate along bicharacteristic curves (or hypersurfaces). As in the case of the Burger equation, singularities will occur for more or less any smooth data. Thus, a classical, continuously differentiable solution can only be constructed locally in time. Both Hamilton-Jacobi equations and Hamiltonian systems play a fundamental role in classical mechanics as well as in the theory of the propagation of singularities in linear PDEs. The deep connection between Hamiltonian systems and first-order Hamilton-Jacobi equations played an important role in the introduction of the Schrödinger equation into quantum mechanics.

2.2 The Initial-Value Problem for ODEs

Before we can continue with our general presentation of PDEs we need first to discuss, for the sake of comparison, the IVP for ODEs. Let us start with a first-order ODE

$$\partial_x u(x) = f(x, u(x)) \quad (26)$$

subject to the initial condition

$$u(x_0) = u_0. \quad (27)$$

Let us also assume for simplicity that (26) is a scalar equation and that f is a well-behaved function of x and u , such as $f(x, u) = u^3 - u + 1 + \sin x$. From the initial data u_0 we can determine $\partial_x u(x_0)$ by substituting x_0 into (26). If we now differentiate the equation (26) with respect to x and apply the chain rule, we derive the equation

$$\partial_x^2 u(x) = \partial_x f(x, u(x)) + \partial_u f(x, u(x)) \partial_x u(x),$$

which for the example just defined works out to be $\cos x + 3u^2(x) \partial_x u(x) - \partial_x u(x)$. Hence,

$$\partial_x^2 u(x_0) = \partial_x f(x_0, u_0) + \partial_u f(x_0, u_0) \partial_x u_0,$$

and since $\partial_x u(x_0)$ has already been determined we find that $\partial_x^2 u(x_0)$ can also be explicitly calculated from the initial data u_0 . This calculation also involves the function f and its first partial derivatives. Taking higher derivatives of the equation (26) we can recursively determine $\partial_x^3 u(x_0)$, as well as all other higher derivatives of u at x_0 . Therefore, one can in principle determine $u(x)$ with the help of the Taylor series

$$\begin{aligned} u(x) &= \sum_{k \geq 0} \frac{1}{k!} \partial_x^k u(x_0) (x - x_0)^k \\ &= u(x_0) + \partial_x u(x_0) (x - x_0) \\ &\quad + \frac{1}{2!} \partial_x^2 u(x_0) (x - x_0)^2 + \cdots \end{aligned}$$

We say “in principle” because there is no guarantee that the series converges. There is, however, a very important theorem, called the *Cauchy-Kowalewski theorem*, which asserts that if the function f is real analytic, as is certainly the case for our function $f(x, u) = u^3 - u + 1 + \sin x$, then there will be some neighborhood J of x_0 where the Taylor series converges to a real-analytic solution u of the equation. It is then easy to show that the solution thus obtained is the unique solution to (26) that satisfies the initial condition (27). To summarize: if f is a well-behaved function, then the initial-value problem for ODEs has a solution, at least in some time interval, and that solution is unique.

The same result does not always hold if we consider a more general equation of the form

$$a(x, u(x)) \partial_x u = f(x, u(x)), \quad u(x_0) = u_0. \quad (28)$$

Indeed, the recursive argument outlined above breaks down in the case of the scalar equation $(x - x_0) \partial_x u = f(x, u)$ for the simple reason that we cannot even determine $\partial_x u(x_0)$ from the initial condition $u(x_0) =$

u_0 . A similar problem occurs for the equation $(u - u_0)\partial_x u = f(x, u)$. An obvious condition that allows us to extend our previous recursive argument to (28) is to insist that $a(x_0, u_0) \neq 0$. Otherwise, we say that the IVP (28) is *characteristic*. If both a and f are also real analytic, the Cauchy-Kowalewski theorem applies again and we obtain a unique, real-analytic solution of (28) in a small neighborhood of x_0 . In the case of an $N \times N$ system,

$$A(x, u(x))\partial_x u = F(x, u(x)), \quad u(x_0) = u_0,$$

$A = A(x, u)$ is an $N \times N$ matrix, and the *noncharacteristic condition* becomes

$$\det A(x_0, u_0) \neq 0. \quad (29)$$

It turns out, and this is extremely important in the development of the theory of ODEs, that, while the nondegeneracy condition (29) is essential to obtain a unique solution of the equation, the analyticity condition is not at all important: it can be replaced by a simple *local Lipschitz condition* for A and F . It suffices to assume, for example, that their first partial derivatives exist and that they are locally bounded. This is always the case if the first derivatives of A and F are continuous.

Theorem (the fundamental theorem of ODEs). *If the matrix $A(x_0, u_0)$ is invertible and if A and F are continuous and have locally bounded first derivatives, then there is some time interval $J \subset \mathbb{R}$ that contains x_0 , and a unique solution² u defined on J that satisfies the initial conditions $u(x_0) = u_0$.*

The proof of the theorem is based on the *Picard iteration method*. The idea is to construct a sequence of approximate solutions $u_{(n)}(x)$ that converge to the desired solution. Without loss of generality we can assume A to be the identity matrix.³ One starts by setting $u_{(0)}(x) = u_0$ and then defines, recursively,

$$\partial_x u_{(n)}(x) = F(x, u_{(n-1)}(x)), \quad u_{(n-1)}(x_0) = u_0.$$

Observe that at every stage all we need to solve is a very simple linear problem, which makes Picard iteration easy to implement numerically. As we shall see below, variations of this method are also used for solving nonlinear PDEs.

Remark. In general, the local existence theorem is sharp, in the sense that its conditions cannot be

relaxed. We have seen that the invertibility condition for $A(x_0, u_0)$ is necessary. Also, it is not always possible to extend the interval J in which the solution exists to the whole of the real line. As an example, consider the nonlinear equation $\partial_x u = u^2$ with initial data $u = u_0$ at $x = 0$, for which the solution $u = u_0/(1 - xu_0)$ becomes infinite in finite time: in the terminology of PDEs, it *blows up*.

In view of the fundamental theorem and the example mentioned above, one can define the main goals of the mathematical theory of ODEs as follows.

- (i) Find criteria for global existence. In the case of blow-up describe the limiting behavior.
- (ii) In the case of global existence describe the asymptotic behavior of solutions and families of solutions.

Though it is impossible to develop a general theory that achieves both goals (in practice one is forced to restrict oneself to special classes of equations motivated by applications), the general local existence and uniqueness theorem mentioned above provides a powerful unifying theme. It would be very helpful if a similar situation were to hold for general PDEs.

2.3 The Initial-Value Problem for PDEs

In the one-dimensional situation one specifies initial conditions at a point. The natural higher-dimensional analogue is to specify them on hypersurfaces $\mathcal{H} \subset \mathbb{R}^d$, that is, $(d - 1)$ -dimensional subsets (or, to be more precise, submanifolds). For a general equation of order k , that is, one that involves k derivatives, we need to specify the values of u and of its first $k - 1$ derivatives in the direction normal to \mathcal{H} . For example, in the case of the second-order wave equation (3) and the initial hyperplane $t = 0$ we need to specify initial data for u and $\partial_t u$.

If we wish to use initial data of this kind to start obtaining a solution, it is important that the data should not be degenerate. (We have already seen this in the case of ODEs.) For this reason, we make the following general definition.

Definition. Suppose that we have a k th-order quasi-linear system of equations, and the initial data comes in the form of the first $k - 1$ normal derivatives that a solution u must satisfy on a hypersurface \mathcal{H} . We say that the system is *noncharacteristic* at a point x_0 of \mathcal{H} if we can use the initial data to determine formally all

2. Since we are not assuming that A and F are analytic, the solution may not be analytic, but it does have continuous first derivatives.

3. Since A is invertible we can multiply both sides of the equation by the inverse matrix A^{-1} .

the other higher partial derivatives of u at x_0 , in terms of the data.

As a very rough picture to have in mind, it may be helpful to imagine an infinitesimally small neighborhood of x_0 . If the hypersurface \mathcal{H} is smooth, then its intersection with this neighborhood will be a piece of a $(d-1)$ -dimensional affine subspace. The values of u and the first $k-1$ normal derivatives on this intersection are given by the initial data, and the problem of determining the other partial derivatives is a problem in linear algebra (because everything is infinitesimally small). To say that the system is noncharacteristic at x_0 is to say that this linear algebra problem can be uniquely solved, which is the case provided that a certain matrix is invertible. This is the nondegeneracy condition referred to earlier.

To illustrate the idea, let us look at first-order equations in two space dimensions. In this case \mathcal{H} is a curve Γ , and since $k-1=0$ we must specify the restriction of u to $\Gamma \subset \mathbb{R}^2$ but we do not have to worry about any derivatives. Thus, we are trying to solve the system

$$\begin{aligned} a^1(x, u(x))\partial_1 u(x) + a^2(x, u(x))\partial_2 u(x) \\ = f(x, u(x)), \quad u|_\Gamma = u_0, \end{aligned} \quad (30)$$

where a^1 , a^2 , and f are real-valued functions of x (which belongs to \mathbb{R}^2) and u . Assume that in a small neighborhood of a point p the curve Γ is described parametrically as the set of points $x = (x^1(s), x^2(s))$. We denote by $n(s) = (n_1(s), n_2(s))$ a unit normal to Γ .

As in the case of ODEs, which we looked at earlier, we would like to find conditions on Γ such that for a given point in Γ we can determine all derivatives of u from the data u_0 , the derivatives of u along Γ , and the equation (30). Out of all possible curves Γ we distinguish in particular the *characteristic* ones we have already encountered above (see (20)):

$$\left. \begin{aligned} \frac{dx^1}{ds} &= a^1(x(s), u(x(s))), \\ \frac{dx^2}{ds} &= a^2(x(s), u(x(s))), \end{aligned} \right\} \quad x(0) = p.$$

One can prove the following fact:

Along a characteristic curve, the equation (30) is degenerate. That is, we cannot determine the first-order derivatives of u uniquely in terms of the data u_0 .

In terms of the rough picture above, at each point there is a direction such that if the hypersurface, which in this case is a line, is along that direction, then the

resulting matrix is singular. If you follow this direction, then you travel along a characteristic curve.

Conversely, if the nondegeneracy condition

$$a^1(p, u(p))n_1(p) + a^2(p, u(p))n_2(p) \neq 0 \quad (31)$$

is satisfied at some point $p = x(0) \in \Gamma$, then we can determine all higher derivatives of u at x_0 uniquely in terms of the data u_0 and its derivatives along Γ . If the curve Γ is given by the equation $\psi(x^1, x^2) = 0$, with nonvanishing gradient $D\psi(p) \neq 0$, then the condition (31) takes the form

$$a^1(p, u(p))\partial_1 \psi(p) + a^2(p, u(p))\partial_2 \psi(p) \neq 0.$$

With a little more work one can extend the above discussion to higher-order equations in higher dimensions, and even to systems of equations. Particularly important is the case of a second-order scalar equation in \mathbb{R}^d ,

$$\sum_{i,j=1}^d a^{ij}(x)\partial_i \partial_j u = f(x, u(x)), \quad (32)$$

together with a hypersurface \mathcal{H} in \mathbb{R}^d defined by the equation $\psi(x) = 0$, where ψ is a function with nonvanishing gradient $D\psi$. Define the unit normal at a point $x_0 \in \mathcal{H}$ to be $n = D\psi/|D\psi|$, or, in component form, $n_i = \partial_i \psi/|\partial \psi|$. As initial conditions for (32) we prescribe the values of u and its normal derivative $n[u](x) = n_1(x)\partial_1 u(x) + n_2(x)\partial_2 u(x) + \dots + n_d(x)\partial_d u(x)$ on \mathcal{H} :

$$u(x) = u_0(x), \quad n[u](x) = u_1(x), \quad x \in \mathcal{H}.$$

It can be shown that \mathcal{H} is noncharacteristic (with respect to equation (32)) at a point p (that is, we can determine all derivatives of u at p in terms of the initial data u_0, u_1) if and only if

$$\sum_{i,j=1}^d a^{ij}(p)\partial_i \psi(p)\partial_j \psi(p) \neq 0. \quad (33)$$

On the other hand, \mathcal{H} is a characteristic hypersurface for (32) if

$$\sum_{i,j=1}^d a^{ij}(x)\partial_i \psi(x)\partial_j \psi(x) = 0 \quad (34)$$

for every x in \mathcal{H} .

Example. If the coefficients a of (32) satisfy the condition

$$\sum_{i,j=1}^d a^{ij}(x)\xi_i \xi_j > 0, \quad \forall \xi \in \mathbb{R}^d, \quad \forall x \in \mathbb{R}^d, \quad (35)$$

then clearly, by (34), no surface in \mathbb{R}^d can be characteristic. This is the case, in particular, for the Laplace

equation $\Delta u = f$. Consider also the minimal-surface equation (7) written in the form

$$\sum_{i,j=1,2} h^{ij}(\partial u) \partial_i \partial_j u = 0, \quad (36)$$

with $h^{11}(\partial u) = 1 + (\partial_2 u)^2$, $h^{22}(\partial u) = 1 + (\partial_1 u)^2$, $h^{12}(\partial u) = h^{21}(\partial u) = -\partial_1 u \partial_2 u$. It is easy to check that the quadratic form associated with the symmetric matrix $h^{ij}(\partial u)$ is positive definite for every ∂u . Indeed,

$$h^{ij}(\partial u) \xi_i \xi_j = (1 + |\partial u|^2)^{-1/2} (|\xi|^2 - (1 + |\partial u|^2)^{-1} (\xi \cdot \partial u)^2) > 0.$$

Thus, even though (36) is not linear, we see that all surfaces in \mathbb{R}^2 are noncharacteristic.

Example. Consider the wave equation $\square u = f$ in \mathbb{R}^{1+d} . All hypersurfaces of the form $\psi(t, x) = 0$ for which

$$(\partial_t \psi)^2 = \sum_{i=1}^d (\partial_i \psi)^2 \quad (37)$$

are characteristic. This is the famous eikonal equation, which plays a fundamental role in the study of wave propagation. Observe that it splits into two Hamilton-Jacobi equations (see (24)):

$$\partial_t \psi = \pm \left(\sum_{i=1}^d (\partial_i \psi)^2 \right)^{1/2}. \quad (38)$$

The bicharacteristic curves of the associated Hamiltonians are called bicharacteristic curves of the wave equation. As particular solutions of (37) we find $\psi_+(t, x) = (t - t_0) + |x - x_0|$ and $\psi_-(t, x) = (t - t_0) - |x - x_0|$, whose level surfaces $\psi_{\pm} = 0$ correspond to forward and backward light cones with their vertex at $p = (t_0, x_0)$. These represent, physically, the union of *all light rays emanating from a point source at p*. The light rays are given by the equation $(t - t_0)\omega = (x - x_0)$, for $\omega \in \mathbb{R}^3$ with $|\omega| = 1$, and are precisely the (t, x) components of the bicharacteristic curves of the Hamilton-Jacobi equations (38). More generally, the characteristics of the linear wave equation

$$a^{00}(t, x) \partial_t^2 u - \sum_{i,j} a^{ij}(t, x) \partial_i \partial_j u = 0, \quad (39)$$

with $a^{00} > 0$ and a^{ij} satisfying (35), are given by the Hamilton-Jacobi equations:

$$-a^{00}(t, x) (\partial_t \psi)^2 + a^{ij}(x) \partial_i \psi \partial_j \psi = 0$$

or, equivalently,

$$\partial_t \psi = \pm \left((a^{00})^{-1} \sum_{i,j} a^{ij}(x) \partial_i \psi \partial_j \psi \right)^{1/2}. \quad (40)$$

The bicharacteristics of the corresponding Hamiltonian systems are called bicharacteristic curves of (39).

Remark. In the case of the first-order scalar equations (17) we have seen how knowledge of characteristics can be used to find, implicitly, general solutions. We have also seen that singularities propagate only along characteristics. In the case of second-order equations the characteristics are not sufficient to solve the equations, but they continue to provide important information, such as how the singularities propagate. For example, in the case of the wave equation $\square u = 0$ with smooth initial data u_0, u_1 everywhere except at a point $p = (t_0, x_0)$, the solution u has singularities present at all points of the light cone $-(t - t_0)^2 + |x - x_0|^2 = 0$ with vertex at p . A more refined version of this fact shows that the singularities propagate along bicharacteristics. The general principle here is that *singularities propagate along characteristic hypersurfaces of a PDE*. Since this is a very important principle, it pays to give it a more precise formulation that extends to general boundary conditions, such as the Dirichlet condition for (1).

Propagation of singularities. *If the boundary conditions or the coefficients of a PDE are singular at some point p, and otherwise smooth (or real analytic) everywhere in some small neighborhood V of p, then a solution of the equation cannot be singular in V except along a characteristic hypersurface passing through p. In particular, if there are no such characteristic hypersurfaces, then any solution of the equation must be smooth (or real analytic) at every point of V other than p.*

Remarks. (i) The heuristic principle mentioned above is invalid, in general, at large scales. Indeed, as we have shown in the case of the Burger equation, solutions to nonlinear evolution equations can develop new singularities whatever the smoothness of the initial conditions. Global versions of the principle can be formulated for linear equations based on the bicharacteristics of the equation. See (iii) below.

(ii) According to the principle, it follows that any solution of the equation $\Delta u = f$, satisfying the boundary condition $u|_{\partial D} = u_0$ with a boundary value u_0 that merely has to be continuous, is automatically smooth everywhere in the interior of D provided that f itself is smooth there. Moreover, the solution is real analytic if f is real analytic.

(iii) More precise versions of this principle, which plays a fundamental role in the general theory, can be given for linear equations. In the case of the general wave

equation (39), for example, one can show that singularities propagate along bicharacteristics. These are the bicharacteristic curves associated with the Hamilton–Jacobi equation (40).

2.4 The Cauchy–Kowalewski Theorem

In the case of ODEs we have seen that a noncharacteristic IVP always admits solutions locally (that is, in some time interval about a given point). Is there a higher-dimensional analogue of this fact? The answer is yes, provided that we restrict ourselves to the real-analytic situation, which is covered by an appropriate extension of the Cauchy–Kowalewski theorem. More precisely, one can consider general quasilinear equations, or systems, with real-analytic coefficients, real-analytic hypersurfaces \mathcal{H} , and appropriate real-analytic initial data on \mathcal{H} .

Theorem (Cauchy–Kowalewski (CK)). *If all the real-analyticity conditions made above are satisfied and if the initial hypersurface \mathcal{H} is noncharacteristic at x_0 ,⁴ then in some neighborhood of x_0 there is a unique real-analytic solution $u(x)$ that satisfies the system of equations and the corresponding initial conditions.*

In the special case of linear equations, an important companion theorem, due to Holmgren, asserts that the analytic solution given by the CK theorem is unique in the class of all smooth solutions and smooth noncharacteristic hypersurfaces \mathcal{H} . The CK theorem shows that, given the noncharacteristic condition and the analyticity assumptions, the following straightforward way of finding solutions works: look for a formal expansion of the kind $u(x) = \sum_{\alpha} C_{\alpha}(x - x_0)^{\alpha}$ by determining the constants C_{α} recursively from simple algebraic formulas arising from the equation and initial conditions on \mathcal{H} . More precisely, the theorem ensures that the naive expansion obtained in this way converges in a small neighborhood of $x_0 \in \mathcal{H}$.

It turns out, however, that the analyticity conditions required by the CK theorem are much too restrictive, and therefore the apparent generality of the result is misleading. A first limitation becomes immediately apparent when we consider the wave equation $\square u = 0$. A fundamental feature of this equation is *finite speed of propagation*, which means, roughly speaking, that if at some time t a solution u is zero outside some bounded set, then the same must be true at all later times.

However, analytic functions cannot have this property unless they are identically zero (see SOME FUNDAMENTAL MATHEMATICAL DEFINITIONS [I.3 §5.6]). Therefore, it is impossible to discuss the wave equation properly within the class of real-analytic solutions. A related problem, first pointed out by HADAMARD [VI.65], concerns the impossibility of solving the Cauchy problem, in many important cases, for arbitrary smooth nonanalytic data. Consider, for example, the Laplace equation $\Delta u = 0$ in \mathbb{R}^d . As we have established above, any hypersurface \mathcal{H} is noncharacteristic, yet the Cauchy problem $u|_{\mathcal{H}} = u_0$, $n[u]|_{\mathcal{H}} = u_1$, for arbitrary smooth initial conditions u_0, u_1 , may admit no local solutions in a neighborhood of any point of \mathcal{H} . Indeed, take \mathcal{H} to be the hyperplane $x_1 = 0$ and assume that the Cauchy problem can be solved for given nonanalytic smooth data in a domain that includes a closed ball B centered at the origin. The corresponding solution can also be interpreted as the solution to the Dirichlet problem in B , with the values of u prescribed on the boundary ∂B . But this, according to our heuristic principle (which can easily be made rigorous in this case), must be real analytic everywhere in the interior of B , contradicting our assumptions about the initial data.

On the other hand, the Cauchy problem for the wave equation $\square u = 0$ in \mathbb{R}^{d+1} has a unique solution for any smooth initial data u_0, u_1 that is prescribed on a *spacelike hypersurface*. This means a hypersurface $\psi(t, x) = 0$ such that at every point $p = (t_0, x_0)$ that belongs to it the normal vector at p lies inside the light cone (either in the future direction or in the past direction). To say this analytically,

$$|\partial_t \psi(p)| > \left(\sum_{i=1}^d |\partial_i \psi(p)|^2 \right)^{1/2}. \quad (41)$$

This condition is clearly satisfied by a hyperplane of the form $t = t_0$, but any other hypersurface close to this is also spacelike. By contrast, the IVP is *ill-posed* for a timelike hypersurface, i.e., a hypersurface for which

$$|\partial_t \psi(p)| < \left(\sum_{i=1}^d |\partial_i \psi(p)|^2 \right)^{1/2}.$$

That is, we cannot, for general non-real-analytic initial conditions, find a solution of the IVP. An example of a timelike hypersurface is given by the hyperplane $x^1 = 0$. Let us explain the term “ill-posed” more precisely.

Definition. A given problem for a PDE is said to be well-posed if both existence and uniqueness of solutions can be established for arbitrary data that belongs to a specified large space of functions, which includes

4. For second-order equations of the kind of (32), this is precisely condition (33).

the class of smooth functions.⁵ Moreover, the solutions must depend continuously on the data. A problem that is not well-posed is called ill-posed.

The continuous dependence on the data is very important. Indeed, the IVP would be of little use if very small changes in the initial conditions resulted in very large changes in the corresponding solutions.

2.5 Standard Classification

The different behavior of the Laplace and wave equations mentioned above illustrates the fundamental difference between ODEs and PDEs and the illusory generality of the CK theorem. Given that these two equations are so important in geometric and physical applications, it is of great interest to find the broadest classes of equations with which they share their main properties. The equations modeled by the Laplace equation are called *elliptic*, while those modeled by the wave equation are called *hyperbolic*. The other two important models are the heat equation (see (2)) and the Schrödinger equation (see (4)). The general classes of equations that they resemble are called *parabolic* and *dispersive*, respectively.

Elliptic equations are the most robust and the easiest to characterize: they are the ones that admit no characteristic hypersurfaces.

Definition. A linear, or quasilinear, $N \times N$ system with no characteristic hypersurfaces is called elliptic.

Equations of type (32) whose coefficients a^{ij} satisfy condition (35) are clearly elliptic. The minimal-surface equation (7) is also elliptic. It is also easy to verify that the Cauchy–Riemann system (12) is elliptic. As was pointed out by Hadamard, the IVP is not well-posed for elliptic equations. The natural way of parametrizing the set of solutions to an elliptic PDE is to prescribe conditions for u , and some of its derivatives (the number of derivatives will be roughly half the order of the equation) at the boundary of a domain $D \subset \mathbb{R}^n$. These are called *boundary-value problems (BVPs)*. A typical example is the Dirichlet boundary condition $u|_{\partial D} = u_0$ for the Laplace equation $\Delta u = 0$ in a domain $D \subset \mathbb{R}^n$. One can show that, if the domain D satisfies certain mild regularity assumptions and the boundary value u_0 is continuous, then this problem admits a unique solution that depends continuously on u_0 . We say that

the Dirichlet problem for the Laplace equation is well-posed. Another well-posed problem for the Laplace equation is given by the Neumann boundary condition $n[u]|_{\partial D} = f$, where n is the exterior unit normal to the boundary. This problem is well-posed for all continuous functions f defined on ∂D with zero mean average. A typical problem of general theory is to classify all well-posed BVPs for a given elliptic system.

As a consequence of our propagation-of-singularities principle, we deduce, heuristically at least, the following general fact:

*Classical solutions of elliptic equations with smooth (or real-analytic) coefficients in a regular domain D are smooth (or real analytic) in the interior of D , whatever the degree of smoothness of the boundary conditions.*⁶

Hyperbolic equations are, essentially, those for which the IVP is well-posed. In that sense, they provide the natural class of equations for which one can prove a result similar to the local existence theorem for ODEs. More precisely, for each sufficiently regular set of initial conditions there is a unique solution. We can thus think of the Cauchy problem as a natural way of parametrizing the set of all solutions to the equations.

The definition of hyperbolicity depends, however, on the particular hypersurface we are considering as the initial hypersurface. Thus, in the case of the wave equation $\square u = 0$, the standard IVP

$$u(0, x) = u_0(x), \quad \partial_t u(0, x) = u_1$$

is well-posed. This means that for any smooth initial data u_0, u_1 we can find a unique solution of the equation, which depends continuously on u_0, u_1 . As we have already mentioned, the IVP for $\square u = 0$ remains well-posed if we replace the initial hypersurface $t = 0$ by any spacelike hypersurface $\psi(t, x) = 0$ (see (41)). However, it fails to be well-posed for timelike hypersurfaces, for which there may be no solution with prescribed, nonanalytic, Cauchy data.

It is more difficult to give algebraic conditions for hyperbolicity. Roughly speaking, hyperbolic equations are at the opposite end of the spectrum from elliptic equations: whereas elliptic equations have no characteristic hypersurfaces, hyperbolic equations have as many as possible passing through any given point. One of the most useful classes of hyperbolic equations,

5. Here we are necessarily vague. A precise space can be specified in each given case.

6. Provided that the boundary condition under consideration is well-posed. Moreover, this heuristic principle holds, in general, only for classical solutions of a nonlinear equation. There are in fact examples of well-posed BVPs, for certain nonlinear elliptic systems, with no classical solutions.

which includes most of the important known examples, consists of equations of the form

$$A^0(t, x, u) \partial_t u + \sum_{i=1}^d A_i(t, x, u) \partial_i u = F(t, x, u), \quad u|_{\mathcal{H}} = u_0, \quad (42)$$

where all the coefficients A^0, A^1, \dots, A^d are symmetric $N \times N$ matrices and \mathcal{H} is given by $\psi(t, x) = 0$. Such a system is well-posed provided that the matrix

$$A^0(t, x, u) \partial_t \psi(t, x) + \sum_{i=1}^d A_i(t, x, u) \partial_i \psi(t, x) \quad (43)$$

is positive definite. A system (42) that satisfies these conditions is called *symmetric hyperbolic*. In the particular case when $\psi(t, x) = t$, the condition (43) becomes

$$(A^0 \xi, \xi) \geq c |\xi|^2 \quad \forall \xi \in \mathbb{R}^N.$$

The following is a fundamental result in the theory of general hyperbolic equations. It is called the local existence and uniqueness of solutions for symmetric hyperbolic systems.

Theorem (fundamental theorem for hyperbolic equations). *The IVP (42) is locally well-posed for symmetric hyperbolic systems with sufficiently smooth A, F , and \mathcal{H} and sufficiently smooth initial conditions u_0 . In other words, if the appropriate smoothness conditions are satisfied, then for any point $p \in \mathcal{H}$ there is a small neighborhood \mathcal{D} of p inside which there is a unique, continuously differentiable solution u .*

Remarks. (i) The local character of the theorem is essential, just as it was for the general propagation-of-singularities principle discussed earlier, since the result cannot be globalized in the particular case of the Burger equation (21), which fits trivially into the framework of general nonlinear symmetric hyperbolic systems. A precise version of the theorem above gives a lower bound on how large \mathcal{D} can be.

(ii) The proof of the theorem is based on a variation of the Picard iteration method that we encountered earlier for ODEs. One starts by taking $u_{(0)} = u_0$ in a neighborhood of \mathcal{H} . Then one defines functions $u_{(n)}$ recursively as follows:

$$\begin{aligned} A^0(t, x, u_{(n-1)}) \partial_t u_{(n)} + \sum_{i=1}^d A_i(t, x, u_{(n-1)}) \partial_i u_{(n)} \\ = F(t, x, u_{(n-1)}), \quad u_{(n)}|_{\mathcal{H}} = u_0. \end{aligned}$$

7. By “point” we mean that p is a spacetime point $(t, x) \in \mathbb{R}^{1+d}$. Similarly, \mathcal{D} is a set of spacetime points.

Notice that at each stage of the iteration we have to solve a linear equation. Linearization is an extremely important tool in studying nonlinear PDEs. We can almost never understand their behavior without linearizing them around important special solutions. Thus, almost invariably, hard problems in nonlinear PDEs reduce to understanding specific problems in linear PDEs.

(iii) To implement the Picard iteration method we need to get precise estimates concerning $u_{(n)}$ in terms of $u_{(n-1)}$. This step requires *energy type a priori estimates*, which we will discuss in section 3.3.

Another important property of hyperbolic equations (which is not shared by elliptic, parabolic, or dispersive equations) is *finite speed of propagation*, which was mentioned earlier in the case of the wave equation (3). Consider this simple case again. The IVP can be solved explicitly by the so-called *Kirchhoff formula*. The formula allows us to conclude that if the initial data at $t = 0$ is zero outside a ball $B_a(x_0)$ of radius $a > 0$ centered at $x_0 \in \mathbb{R}^3$, then at time $t > 0$ the solution u is zero outside the ball $B_{a+t}(x_0)$. In general, finite speed of propagation can best be formulated in terms of domains of dependence and influence of hyperbolic equations (see the online version for general definitions).

Hyperbolic PDEs play a fundamental role in physics, as they are intimately tied to the relativistic nature of the modern theory of fields. Equations (3), (5), (13) are the simplest examples of *linear field theories*, and they are manifestly hyperbolic. Other basic examples appear in *gauge field theories* such as MAXWELL’S EQUATIONS [IV.13 §1.1] $\partial^\alpha F_{\alpha\beta} = 0$ or the *Yang-Mills equations* $D^\alpha F_{\alpha\beta} = 0$. Finally, the Einstein equations (14) are also hyperbolic.⁸ Other important examples of hyperbolic equations arise in the physics of elasticity and inviscid fluids. As examples of the latter, the Burger equation (21) and the compressible Euler equation are hyperbolic.

Elliptic equations, on the other hand, appear naturally in describing time-independent, or more generally *steady-state*, solutions of hyperbolic equations. Elliptic equations can also be derived, directly, by well-defined VARIATIONAL PRINCIPLES [III.96].

Finally, a few words about parabolic equations and Schrödinger-type equations, which are intermediate

8. For gauge theories and Einstein equations the notion of hyperbolicity depends on the choice of gauge or coordinates. In the case of the Yang-Mills equations, for example, one obtains a well-defined system of nonlinear wave equations only in the Lorentz gauge.

between the elliptic and hyperbolic ones. Large classes of useful equations of these types are given by

$$\partial_t u - Lu = f \quad (44)$$

and

$$i\partial_t u + Lu = f, \quad (45)$$

respectively, where L is an elliptic second-order operator. One looks for solutions $u = u(t, x)$, defined for $t \geq t_0$, with the prescribed initial condition

$$u(t_0, x) = u_0(x) \quad (46)$$

on the hypersurface $t = t_0$. Strictly speaking, this hypersurface is characteristic, since the order of the equation is 2 and we cannot determine $\partial_t^2 u$ at $t = t_0$ directly from the equation. Yet this is not a serious problem; we can still determine $\partial_t^2 u$ formally by differentiating the equation with respect to ∂_t . Thus, the IVP (44) (or (45)) with initial condition (46) is well-posed, but not quite in the same sense as for hyperbolic equations. For example, the heat equation $-\partial_t u + \Delta u$ is well-posed for positive t but ill-posed for negative t . The heat equation may also not have unique solutions for the IVP unless we make assumptions about how fast the initial data is allowed to grow at infinity. One can also show that the characteristic hypersurfaces of the equation (44) are all of the form, and therefore parabolic equations are quite similar to elliptic equations. For example, one can show that if the coefficients a^{ij} and f are smooth (or real analytic), then the solution u must be smooth (or real analytic in x) for $t > t_0$ even if the initial data u_0 is not smooth, which is consistent with our propagation-of-singularities principle. The heat equation smooths out initial conditions. It is for this reason that the heat equation is useful in many applications. In physics, parabolic PDEs arise whenever diffusion or dissipation phenomena are important, while in geometry and calculus of variations, parabolic PDEs often arise as gradient flows of positive-definite functionals. Ricci flow (16) can also be viewed as a parabolic PDE, after a suitable change of coordinates.

Dispersive PDEs, of which the Schrödinger equation (4) is a fundamental example, are evolution equations that behave analogously to hyperbolic PDEs in many respects. For instance, the IVP tends to be locally well-posed both forward and backward in time. However, solutions to dispersive PDEs do not propagate along characteristic surfaces. Instead, they move at speeds that are determined by their spatial frequency; in general, high-frequency waves tend to propagate at much

greater speeds than low-frequency waves, which eventually leads to a *dispersion* of the solution into increasingly large areas of space. In fact, the speed of propagation of solutions is typically infinite. This behavior also differs from that of parabolic equations, which tend to *dissipate* the high-frequency components of a solution (sending them to zero) rather than dispersing them. In physics, dispersive equations arise in quantum mechanics: they are the *nonrelativistic limit* $c \rightarrow \infty$ of relativistic equations and they are also approximations to model certain types of fluid behavior. For instance, the KORTEWEG-DE VRIES EQUATION [III.51],

$$\partial_t u + \partial_x^3 u = 6u\partial_x u,$$

is a dispersive PDE that models the behavior of small-amplitude waves in a shallow canal.

2.6 Special Topics for Linear Equations

The greatest successes of the general theory have been in connection with linear equations, especially those with constant coefficients, for which Fourier analysis provides an extremely powerful tool. While the related issues of classification, well-posedness, and propagation of singularities have dominated the study of linear equations, there are other issues of interest as well, including the following.

2.6.1 Local Solvability

This is the problem of determining the conditions on a linear operator \mathcal{P} and given data f under which the equation (9) is locally solvable. The Cauchy-Kowalewski theorem gives a criterion for local solvability when f and the coefficients of \mathcal{P} are real analytic, but it is a remarkable phenomenon that when one relaxes this assumption slightly, asking for f to be smooth rather than real analytic, serious obstructions to local solvability appear. For instance, the *Lewy operator*

$$\mathcal{P}[u](t, z) = \frac{\partial u}{\partial \bar{z}}(t, z) - iz \frac{\partial u}{\partial t}(t, z),$$

defined on complex-valued functions $u : \mathbb{R} \times \mathbb{C} \rightarrow \mathbb{C}$, has the property that equation (9) is locally solvable for real-analytic f but not for “most” smooth f . The Lewy operator is intimately connected to the tangential Cauchy-Riemann equations on the Heisenberg group in \mathbb{C}^2 . It was discovered in the study of the restriction of the two-dimensional analogue of the Cauchy-Riemann operator \mathcal{P} to a quadric in \mathbb{C}^2 . This example was the starting point for the theory of *local solvability*, whose goal is to characterize linear equations that are locally

solvable. The theory of Cauchy–Riemann manifolds—which has its origin in the study of restrictions of the Cauchy–Riemann equations (in higher dimensions) to real hypersurfaces, each of which comes with an associated “tangential Cauchy–Riemann complex”—is another extremely rich source of examples of interesting linear PDEs, which do not fit into the standard classification.

2.6.2 Unique Continuation

This concerns various ill-posed problems where solutions may not always exist, but one still has uniqueness. A fundamental example is that of *analytic continuation*: two holomorphic functions on a connected domain D that agree on a nondiscrete set (such as a disk or an interval) must necessarily agree everywhere on D . This fact can be viewed as a unique continuation result for the Cauchy–Riemann equations (12). Another example in a similar spirit is *Holmgren’s theorem*, which asserts that solutions to a linear PDE (9) that has real-analytic coefficients and data are unique, even in the class of smooth functions. More generally, the study of ill-posed problems (such as the wave equation with prescribed data on a timelike surface rather than a space-like one) arises naturally in connection with control theory.

2.6.3 Spectral Theory

There is no way I can even begin to give an account of this theory, which is of fundamental importance not only to quantum mechanics and other physical theories, but also to geometry and ANALYTIC NUMBER THEORY [IV.2]. Just as a matrix A can often be analyzed through its EIGENVALUES AND EIGENVECTORS [I.3 §4.3] by the tools of linear algebra, one can learn much about a linear differential operator \mathcal{P} and its associated PDE by understanding that operator’s SPECTRUM [III.88] and eigenfunctions with the help of tools from FUNCTIONAL ANALYSIS [IV.15]. A typical problem in spectral theory is the *eigenvalue problem* in \mathbb{R}^d :

$$-\Delta u(x) + V(x)u(x) = \lambda u(x).$$

A function u that is localized in space (for example, by being bounded in the $L^2(\mathbb{R}^d)$ -norm) and that satisfies this equation is mapped by the linear operator $-\Delta + V$ to the function λu : we say that u is an *eigenfunction* with *eigenvalue* λ .

Suppose that we have an eigenfunction u and let $\phi(t, x) = e^{-i\lambda t}u(x)$. It is easy to check that ϕ is a

solution of the Schrödinger equation

$$i\partial_t \phi + \Delta \phi - V\phi = 0. \quad (47)$$

Moreover, it has a very special form. Such solutions are called *bound states* of the physical system described by (47). The eigenvalues λ , which form a discrete set, correspond to the quantum energy levels of the system. They are very sensitive to the choice of potential V . The *inverse spectral problem* is also important: can one determine the potential V from knowledge of the corresponding eigenvalues? The eigenvalue problem can be studied in considerable generality by replacing the operator $-\Delta + V$ with other elliptic operators. For instance, in geometry it is important to study the eigenvalue problem for the *Laplace–Beltrami operator*, which is the natural generalization of the Laplace operator from \mathbb{R}^n to general RIEMANNIAN MANIFOLDS [I.3 §6.10]. When the manifold has some arithmetic structure (for instance, if it is the quotient of the upper half-plane by a discrete arithmetic group), this problem is of major importance in number theory, leading, for instance, to the theory of *Hecke–Maas forms*. A famous problem in differential geometry (“can you hear the shape of a drum?”) is to characterize the metric on a compact surface from the spectral properties of the associated Laplace–Beltrami operator.

2.6.4 Scattering Theory

This theory formalizes the intuition from quantum mechanics that a potential which is small or localized is largely unable to “trap” a quantum particle, which is therefore likely to escape to infinity in a manner resembling that of a free particle. In the case of equation (47), solutions that scatter are those that behave freely as $t \rightarrow \infty$. That is, they behave like solutions to the free Schrödinger equation $i\partial_t \psi + \Delta \psi = 0$. A typical problem in scattering theory is to show that, if $V(x)$ tends to zero sufficiently fast as $|x| \rightarrow \infty$, all solutions, except the bound states, scatter as $t \rightarrow \infty$.

2.7 Conclusions

In the analytic case, the CK theorem allows us to solve the IVP locally for very general classes of PDEs. We have a general theory of characteristic hypersurfaces of PDEs and a good general understanding of how they relate to propagation of singularities. We can also distinguish in considerable generality the fundamental classes of elliptic and hyperbolic equations and can define general parabolic and dispersive equations. The IVP for

a large class of nonlinear hyperbolic systems can be solved locally in time, for sufficiently smooth initial conditions. Similar local-in-time results hold for general classes of nonlinear parabolic and dispersive equations. For linear equations a lot more can be done. We have satisfactory results concerning the regularity of solutions for elliptic and parabolic equations and a good understanding of the propagation of singularities for a large class of hyperbolic equations. Some aspects of spectral theory and scattering theory and problems of unique continuation can also be studied in considerable generality.

The main defect of the general theory concerns the passage from local to global. Important global features of special equations are too subtle to fit into a general scheme. Rather, each important PDE requires special treatment. This is particularly true for nonlinear equations: the long-term behavior of solutions is very sensitive to the special features of the equation at hand. Moreover, general points of view may obscure, through unnecessary technical complications, the main properties of the important special cases. A useful general framework is one that provides a simple and elegant treatment of a particular phenomenon, as is the case for symmetric hyperbolic systems and the phenomenon of local well-posedness and finite speed of propagation. However, it turns out that symmetric hyperbolic systems are simply too general for the study of more refined questions about the important examples of hyperbolic equations.

3 General Ideas

As one turns away from the general theory, one may be inclined to accept the pragmatic point of view described earlier, according to which PDEs is not a real subject but is rather a collection of subjects such as hydrodynamics, general relativity, several complex variables, elasticity, etc., each organized around a special equation. However, this rather widespread viewpoint has its own serious drawbacks. Even though specific equations have specific properties, the tools that are used to derive them are intimately related. In fact, there is an impressive body of knowledge relevant to all important equations, or at least large classes of them. Lack of space does not allow me to do anything more than enumerate them below.⁹

9. I fail to mention in the few examples given above some of the important functional analytic tools connected to Hilbert space methods, compactness, the implicit function theorems, etc. I also fail to mention the importance of probabilistic methods and the develop-

3.1 Well-Posedness

As is clear from the previous section, well-posed problems are at the heart of the modern theory of PDEs. Recall that these are problems that admit unique solutions for given smooth initial or boundary conditions, and that the corresponding solutions have to depend continuously on the data. It is this condition that leads to the classification of PDEs into elliptic, hyperbolic, parabolic, and dispersive equations. The first step in the study of a nonlinear evolution equation is a proof of a local-in-time existence and uniqueness theorem, similar to the one for ODEs. *Ill-posedness*, the counterpart of well-posedness, is also important in many applications. The Cauchy problem for the wave equation (3), with data on the timelike hypersurface $z = 0$, is a typical example. Ill-posed problems appear naturally in control theory and inverse scattering.

3.2 Explicit Representations and Fundamental Solutions

Our basic equations (2)–(5) can be solved explicitly. For example, the solution to the IVP for the heat equation in \mathbb{R}_+^{1+d} , that is, the problem of finding a function u that satisfies

$$-\partial_t u + \Delta u = 0, \quad u(0, x) = u_0(x),$$

for $t \geq 0$, is given by

$$u(t, x) = \int_{\mathbb{R}^d} E_d(t, x - y) u_0(y) dy$$

for a certain function E_d , which is called the *fundamental solution* of the heat operator $-\partial_t + \Delta$. This function can be defined explicitly: when $t \leq 0$ it is 0, and when $t > 0$ it is given by the formula $E_d(t, x) = (4\pi t)^{-d/2} e^{-|x|^2/4t}$. Observe that E_d satisfies the equation $(-\partial_t + \Delta)E = 0$ in both regions $t < 0$ and $t > 0$, but it has a singularity at $t = 0$, which prevents it from satisfying the equation in the whole of \mathbb{R}^{1+d} . In fact, we can check that for any function¹⁰ $\phi \in C_0^\infty(\mathbb{R}^{d+1})$, we have

$$\int_{\mathbb{R}^{d+1}} E_d(t, x) (\partial_t \phi(t, x) + \Delta \phi(t, x)) dt dx = \phi(0, 0). \quad (48)$$

In the language of DISTRIBUTION THEORY [III.18], formula (48) means that E_d , as a distribution, satisfies the equation $(-\partial_t + \Delta)E_d = \delta_0$, where δ_0 is the *Dirac*

ment of topological methods for dealing with global properties of elliptic PDEs.

10. That is, any function that is smooth and has compact support in \mathbb{R}^{1+d} .

distribution in \mathbb{R}^{1+d} supported at the origin. That is, $\delta_0(\phi) = \phi(0, 0)$, $\forall \phi \in C_0^\infty(\mathbb{R}^{d+1})$. A similar notion of fundamental solution can be defined for the Poisson, wave, Klein–Gordon, and Schrödinger equations.

A powerful method of solving linear PDEs with constant coefficients is based on THE FOURIER TRANSFORM [III.27]. For example, consider the heat equation $\partial_t u - \Delta u = 0$ in one space dimension, with initial condition $u(0, x) = u_0$. Define $\hat{u}(t, \xi)$ to be the Fourier transform of u relative to the space variable:

$$\hat{u}(t, \xi) = \int_{-\infty}^{+\infty} e^{-ix\xi} u(t, x) dx.$$

It is easy to see that $\hat{u}(t, \xi)$ satisfies the differential equation

$$\partial_t \hat{u}(t, \xi) = -\xi^2 \hat{u}(t, \xi), \quad \hat{u}(0, \xi) = \hat{u}_0(\xi).$$

This can be solved by a simple integration, which results in the formula $\hat{u}(t, \xi) = \hat{u}_0(\xi) e^{-t|\xi|^2}$. Thus, with the help of the inverse Fourier transform, we derive a formula for $u(t, x)$:

$$u(t, x) = (2\pi)^{-1} \int_{-\infty}^{+\infty} e^{ix\xi} e^{-t|\xi|^2} \hat{u}_0(\xi) d\xi.$$

Similar formulas can be derived for our other basic evolution equations. For example, in the case of the wave equation $-\partial_t^2 u + \Delta u = 0$ in three dimensions, subject to the initial data $u(0, x) = u_0$, $\partial_t u(0, x) = 0$, we find that

$$u(t, x) = (2\pi)^{-3} \int_{\mathbb{R}^3} e^{ix\xi} \cos(t|\xi|) \hat{u}_0(\xi) d\xi. \quad (49)$$

After some work, one can reexpress formula (49) in the form

$$u(t, x) = \partial_t \left((4\pi t)^{-1} \int_{|x-y|=t} u_0(y) da(y) \right), \quad (50)$$

where da is the area element of the sphere $|x-y| = t$ of radius t centered at x . This is the well-known *Kirchhoff formula*. By contrast with (49), the integration here is with respect to the physical variables t and x only. It is instructive to compare these two formulas. Using the Plancherel identity it is very easy to deduce from (49) the L^2 bound

$$\int_{\mathbb{R}^3} |u(t, x)|^2 dx \leq C \|u_0\|_{L^2(\mathbb{R}^3)}^2,$$

while the possibility of obtaining such a bound from (50) seems unlikely since the formula involves a derivative. On the other hand, (50) is perfect for giving us information about the domain of influence. Indeed, we can see immediately from the formula that if u_0 is zero outside the ball $B_a = \{|x - x_0| \leq a\}$, then $u(t, x)$ is zero outside the ball $B_{a+|t|}$ for any time t . This fact does not seem at all transparent in the Fourier-based

formula (49). The fact that different representations of solutions have different, even opposite, strengths and weaknesses has important consequences for constructing approximate solutions, or *parametrics*, for more complicated equations, such as linear equations with variable coefficients or nonlinear wave equations. There are two possible types of constructions: those in physical space, which mimic the physical-space formula (50), and those in Fourier space, which mimic the formula (49).

3.3 A Priori Estimates

Most equations cannot be solved explicitly. However, if we are interested in *qualitative* information about a solution, then it is not necessary to derive it from an exact formula. But how else, one might wonder, can we extract such information? A priori estimates are a very important technique for doing this.

The best-known examples are *energy estimates*, the *maximum principle*, and *monotonicity arguments*. The simplest example of the first type is the following identity (which is a very simple example of a so-called *Bochner-type identity*):

$$\int_{\mathbb{R}^d} |\partial^2 u(x)|^2 dx = \int_{\mathbb{R}^d} |\Delta u(x)|^2 dx.$$

The left-hand side is shorthand for

$$\int_{\mathbb{R}^d} \sum_{1 \leq i, j \leq d} |\partial_i \partial_j u(x)|^2 dx$$

and the identity holds for all functions u that are twice continuously differentiable and tend to zero as $|x| \rightarrow \infty$. This formula can be justified fairly simply by integrating by parts. As a consequence of the Bochner identity, we obtain the a priori estimate that if u is a smooth solution to the Poisson equation (6) with square-integrable data f , and if it tends to zero at infinity, then the square integral of its second derivatives is bounded:

$$\int_{\mathbb{R}^d} |\partial^2 u(x)|^2 dx \leq \int_{\mathbb{R}^d} |f(x)|^2 dx < \infty. \quad (51)$$

Thus we obtain the qualitative fact that, on average (in a mean-square sense), u has “two more degrees of regularity” than f .¹¹ This is called an *energy-type estimate* because, in physical situations, the square of

11. A crucial fact, about which one can read more in the online version, is that the L^2 -norms in (51) can be replaced by L^p -norms, $1 < p < \infty$, or Hölder-type norms. The first case corresponds to *Calderon–Zygmund estimates*, while the second corresponds to *Schauder estimates*. Both are extremely important in the study of regularity properties for solutions to second-order elliptic PDEs.

T&T note: need to check whether this is mentioned by Fefferman in his Euler and Navier–Stokes article.

the L^2 -norm can often be interpreted as some type of kinetic energy.

The Bochner identity can be extended to more general Riemannian manifolds than \mathbb{R}^d , although one then picks up some additional lower-order terms involving the curvature of those manifolds. Such identities play a major role in the theory of geometric PDEs on these manifolds.

Energy-type identities and estimates also exist for parabolic, dispersive, and hyperbolic PDEs. For instance, they play a fundamental role in demonstrating the local existence, uniqueness, and finite speed of propagation for hyperbolic PDEs with smooth initial data. Energy estimates become particularly powerful when combined with inequalities such as the *Sobolev embedding inequality*, which allows one to convert the “ L^2 ” information provided by these estimates into pointwise (or “ L^∞ ”) type information (see FUNCTION SPACES [III.29 §§2.4, 3]).

While energy identities and L^2 estimates (which, as in the above example, come from integration by parts) apply to all, or at least major classes of, PDEs, the *maximum principle* can be applied only to elliptic and parabolic PDEs. The following theorem is the simplest manifestation of it. Note that the theorem provides us with important quantitative information about solutions to the Laplace equation even in the absence of any explicit representation for them.

Theorem (maximum principle). *Assume that u is a solution to the Laplace equation (1) on a bounded connected domain $D \in \mathbb{R}^d$ with a smooth boundary ∂D . Assume also that u is continuous on the closure of D and has continuous first and second partial derivatives in the interior of D . Then u must achieve its maximum and minimum values on the boundary. Moreover, if the maximum or minimum is also achieved at an interior point of D , then u must be constant in D .*

The method is very robust and can easily be extended to a large class of second-order elliptic equations. It can also be extended to parabolic equations and systems, and plays a crucial role in, for example, the study of Ricci flow.

Let us briefly mention some other important classes of a priori estimates. The Sobolev inequalities, which are of prime importance in elliptic equations, have several counterparts in linear and nonlinear hyperbolic and dispersive equations, including the *Strichartz estimates* and *bilinear estimates*. In connection with

ill-posed problems and unique continuation, *Carleman estimates* play a fundamental role. Finally, several a priori estimates arising from monotonicity formulas¹²—such as *virial identities*, *Pohozaev identities*, or *Morawetz inequalities*—can be used to establish the breakdown of regularity or the *blow-up* of solutions to some nonlinear equations, and to guarantee global existence and decay of solutions to others.

To summarize, it is not much of an exaggeration to say that a priori estimates play a fundamental role in more or less every aspect of the modern theory of PDEs.

3.4 Bootstrap and Continuity Arguments

The *bootstrap* argument is a method, or rather a powerful general philosophy, to derive a priori estimates for nonlinear equations. According to this philosophy we start by making educated assumptions about the solutions we are trying to describe. These assumptions allow us to think of the original nonlinear problem as a linear one whose coefficients satisfy properties consistent with the assumptions. We may then use linear methods, based on other a priori estimates that we already know, to try to show that the solutions to this linear problem behave as well as we have postulated—in fact, even better. One can characterize this powerful method, which allows us to use linear theory without actually having to linearize the equation, as a *conceptual linearization*. It can also be regarded as a continuity argument relative to some parameter, which might be the natural time parameter of an evolution problem, but it could also be an artificial parameter which we have the freedom to introduce ourselves. This latter situation is typical of applications to nonlinear elliptic equations. In the online version of this article we provide a few examples to illustrate the method in both cases.

3.5 The Method of Generalized Solutions

Since a PDE involves differentiation, it might seem obvious that in any discussion of PDEs we should restrict our attention to differentiable functions. However, it is possible to generalize the notion of differentiation so that it makes sense for a wider class of functions, and even for function-like objects, such as distributions, that are not functions at all. This allows us to make

12. Perhaps the most familiar example of a monotonicity phenomenon is the *second law of thermodynamics* from physics, which asserts that, for many physical systems, the total *entropy* of the system is an increasing function of time.

sense of a PDE in a broader context, and admits the possibility of *generalized solutions*.

The best way to introduce generalized solutions in PDEs and explain why they are important is through the *Dirichlet principle*. This originates in the observation that, out of all functions that are defined on a bounded domain $D \subset \mathbb{R}^d$, that satisfy prescribed Dirichlet boundary condition $u|_{\partial D} = f$, and that live in an appropriate functional space X , the functions u that minimize the Dirichlet integral (or Dirichlet functional)

$$\|u\|_{D^r}^2 = \frac{1}{2} \int_D |\nabla u|^2 = \frac{1}{2} \sum_{i=1}^d \int_D |\partial_i u|^2 \quad (52)$$

are the harmonic functions (that is, solutions of the equation $\Delta u = 0$). It was RIEMANN [VI.49] who first had the idea of trying to use this fact to solve the Dirichlet problem: in order to find a solution u to the problem

$$\Delta u = 0, \quad u|_{\partial D} = u_0, \quad (53)$$

one should find (by some means other than solving the Dirichlet problem) a function u that minimizes the Dirichlet integral while equaling u_0 on ∂D . To do this, one must specify the set by functions, or rather the function space, over which the minimization is taking place. The history of how this choice was made is a fascinating one. A natural choice is $X = C^1(\bar{D})$, the space of continuously differentiable functions on \bar{D} , where the norm of a function v is

$$\|v\|_{C^1(\bar{D})} = \sup_{x \in \bar{D}} (|v(x)| + |\partial v(x)|).$$

In particular, the Dirichlet norm $\|v\|_{D^r}$ is finite when v belongs to this space. In fact, Riemann chose $X = C^2(\bar{D})$ (a similar space but designed for twice continuously differentiable functions). This bold but flawed attempt was followed by a penetrating criticism by WEIERSTRASS [VI.44], who showed that the functional does not have to achieve its minimum in either $C^2(\bar{D})$ or $C^1(\bar{D})$. However, Riemann's basic idea was revived, and it eventually triumphed after a long and inspiring process that involved defining appropriate function spaces, introducing the notion of generalized solutions, and developing a *regularity theory* for them. (The precise formulation of the Dirichlet principle also requires the definition of SOBOLEV SPACES [III.29 §2.4].)

Let us briefly summarize the method, which has since been vastly extended so that it can be applied to a large class of linear¹³ and nonlinear elliptic and parabolic equations. It is based on two steps. In the first step one

applies a minimization procedure. Although, as Weierstrass discovered, the natural function spaces may not contain functions that achieve the minimum, one can use such a procedure to find a *generalized* solution instead. This may not seem very interesting, since we were looking for a *function* that solves the Dirichlet problem (or one of the other problems to which the method can be applied). But this is where the second step comes in: it is sometimes possible to show that the generalized solution must in fact be a classical solution (that is, an appropriately smooth function) after all. This is the “regularity theory” mentioned earlier. In some situations, however, the generalized solution may turn out to have singularities and therefore not be regular. Then the challenge is to understand the nature of these singularities and to prove realistic *partial* regularity results. For instance, it is sometimes possible to prove that the generalized solution is smooth everywhere apart from in a small “exceptional set.”

Though generalized solutions are at their most effective for elliptic problems, their range of applicability encompasses all PDEs. For example, we have already seen that the fundamental solutions to the basic linear equations have to be interpreted as distributions, which are examples of generalized solutions.

The notion of generalized solutions has also proved successful for nonlinear evolution problems, such as systems of conservation laws in one space dimension. An excellent example is provided by the Burger equation (21). As we have seen, solutions to $\partial_t u + u \partial_x u = 0$ develop singularities in finite time no matter how smooth the initial conditions are. It is natural to ask whether solutions continue to make sense, as generalized solutions, even beyond the time when these singularities form. A natural notion of generalized solution is a function u such that

$$\int_{\mathbb{R}^{1+1}} (\partial_t u + u \partial_x u) \phi = 0$$

for every smooth function ϕ that is zero outside a bounded set, since one can make sense of the integral even when u is not a differentiable function. Integrating this by parts (the first term with respect to t and the second with respect to x) one obtains the following formulation:

$$\int_{\mathbb{R}^{1+1}} u \partial_t \phi + \frac{1}{2} \int_{\mathbb{R}^{1+1}} u^2 \partial_x \phi = 0 \quad \forall \phi \in C_0^\infty(\mathbb{R}^{1+1}).$$

It can be shown that, under additional conditions called *entropy conditions*, the IVP for the Burger equation admits a unique generalized solution that is *global*:

13. A notable example for applications in geometry is Hodge theory.

that is, valid for every $t \in \mathbb{R}$. Today we have a satisfactory theory of global solutions to a large class of hyperbolic systems of one-dimensional “conservation laws.” These systems, for which the above-mentioned theory applies, are called *strictly hyperbolic*.

For more complicated nonlinear evolution equations, the question of what constitutes a good concept of a generalized solution, though fundamental, is far murkier. For higher-dimensional evolution equations the first concept of a *weak solution* was introduced by Leray. Let us call a generalized solution *weak* if one cannot prove any type of uniqueness for it. This unsatisfactory situation may be temporary, i.e., the result of our technical inabilities, or unavoidable, in the sense that the concept itself is flawed. Leray was able to produce, by a compactness method, a weak solution of the IVP for the NAVIER-STOKES EQUATIONS [III.23]. The great advantage of the compactness method (and its modern extensions, which can, in some cases, cleverly circumvent lack of compactness) is that it produces global solutions for all data. This is particularly important for supercritical or critical nonlinear evolution equations, which we will discuss later. For these we expect classical solutions to develop singularities in a finite time. The problem, however, is that one has very little control over such solutions. In particular, we do not know how to prove their uniqueness.¹⁴ Similar types of solutions were later introduced for other important nonlinear evolution equations. In most of the interesting cases of supercritical evolution equations, such as the Navier-Stokes equations, the usefulness of the types of weak solutions discovered so far remains undecided.

3.6 Microlocal Analysis, Parametrixes, and Paradifferential Calculus

One of the fundamental difficulties of hyperbolic and dispersive equations is the interplay between geometric properties, which concern the physical space, and other properties, intimately tied to oscillations, that are best seen in Fourier space. *Microlocal analysis* is a general still-developing philosophy according to which one isolates the main difficulties by careful localizations in physical space or Fourier space or both. An important application of this point of view is the construction of parametrixes for linear hyperbolic equations and their use in proving results

about the propagation of singularities. Parametrixes, as we have already mentioned, are approximate solutions of linear equations with variable coefficients, with error terms that are smoother. The *paradifferential calculus* is an extension of microlocal analysis to nonlinear equations. It allows one to manipulate the form of a nonlinear equation by taking account of how large and small frequencies interact, and it has achieved a remarkable technical versatility.

PUP: I can confirm that this statement is accurate.

3.7 Scaling Properties of Nonlinear Equations

A PDE is said to have a *scaling property* if, whenever one rescales a solution in an appropriate way, one obtains another solution. Essentially, all basic nonlinear equations have well-defined scaling properties. Take, for example, the Burger equation (21), $\partial_t u + u \partial_x u = 0$. If u is a solution of this equation, then so is the function u_λ defined by $u_\lambda(t, x) = u(\lambda t, \lambda x)$. Similarly, if u is a solution of the cubic nonlinear Schrödinger equation in \mathbb{R}^d ,

$$i \partial_t u + \Delta u + c |u|^2 u = 0, \quad (54)$$

then so is $u_\lambda(t, x) = \lambda u(\lambda^2 t, \lambda x)$. The relationship between the nonlinear scaling of the equation and the a priori estimates available for solutions to the equations leads to an extremely useful classification of equations into subcritical, critical, and supercritical equations. This will be discussed in more detail in the next section. For the moment it suffices to say that subcritical equations are those for which the nonlinearity can be controlled by the existing a priori estimates of the equation, while supercritical equations are those for which the nonlinearity appears to be stronger. Critical equations are borderline. The definition of criticality and its relationship with the issue of regularity play a very important heuristic role in nonlinear PDEs. One expects supercritical equations to develop singularities and subcritical equations not to.

4 The Main Equations

In the previous section we argued that, while there is no hope of finding a general theory of all PDEs, there is nevertheless a wealth of general ideas and techniques that are relevant to the study of almost all important equations. In this section we indicate how it may be possible to identify the features that characterize the equations we call important.

Most of our basic PDEs can be derived from simple geometric principles, which happen to coincide with

¹⁴. Leray was very concerned about this point. Though, like all other researchers after him, he was unable to prove uniqueness of his weak solution, he managed to show that it must coincide with a classical one as long as the latter does not develop singularities.

some of the underlying geometric principles of modern physics. These simple principles provide a unifying framework¹⁵ for the subject and help endow it with a sense of purpose and cohesion. They also explain why a very small number of linear differential operators, such as the Laplacian and the d'Alembertian, are all-pervasive.

Let us begin with the operators. The Laplacian is the simplest differential operator that is invariant under rigid motions of Euclidean space—a fact that we noted at the beginning of this article. This is important mathematically and physically: mathematically because it results in many symmetry properties and physically because many physical laws are themselves invariant under rigid motions. The d'Alembertian is, similarly, the simplest differential operator that is invariant under the natural symmetries, or Poincaré transformations, of Minkowski space.

Now let us turn to the equations. From the point of view of physics, the heat equation is basic because it is the simplest paradigm for diffusive phenomena, while the Schrödinger equation can be viewed as the Newtonian limit of the Klein-Gordon equation. The geometric framework of the former is Galilean space, which itself is simply the Newtonian limit of Minkowski space.¹⁶

From a mathematical point of view, the heat, Schrödinger, and wave equations are basic because the corresponding differential operators $\partial_t - \Delta$, $(1/i)\partial_t - \Delta$, and $\partial_t^2 - \Delta$ are the simplest evolution operators that can be built out of Δ . The wave operator, as just discussed, is basic in a deeper way because of the association between $\square = -\partial_t^2 + \Delta$ and the geometry of Minkowski space \mathbb{R}^{1+n} . As for Laplace's equation, one can view solutions to $\Delta\phi = 0$ as special time-independent solutions to $\square\phi = 0$. Appropriate invariant and local definitions of square roots of Δ and \square , or $\square - k^2$, corresponding to “spinorial representations” of the Lorentz group, lead to the associated Dirac operators (see (13)). In the same vein we can associate with every Riemannian or Lorentzian manifold the operator Δ_g or \square_g , respectively, or the corresponding Dirac operators. These equations inherit in a straightforward way the symmetries of the spaces on which they are defined.

15. The scheme sketched below is only an attempt to show that, in spite of the enormous number of PDEs studied by mathematicians, physicists, and engineers, there are nevertheless simple basic principles that unite them. I do not want, by any means, to imply that the equations discussed below are the only ones worthy of our attention.

16. This is done by starting with the Minkowski metric $m = \text{diag}(-1/c^2, 1, 1, 1)$, where c corresponds to the velocity of light, and letting $c \rightarrow \infty$.

4.1 Variational Equations

There is a general and extremely effective method for generating equations with prescribed symmetries that plays a fundamental role in both physics and geometry. One starts with a scalar quantity, called a *Lagrangian*, such as

$$\mathcal{L}[\phi] = \sum_{\mu, \nu=0}^3 m^{\mu\nu} \partial_\mu \phi \partial_\nu \phi - V(\phi), \quad (55)$$

with ϕ a real-valued function defined on \mathbb{R}^{1+3} and V some real function of ϕ such as, for example, $V(\phi) = \phi^3$. Here ∂_μ denotes the partial derivatives with respect to the coordinates x^μ , $\mu = 0, 1, 2, 3$, and $m^{\mu\nu} = m_{\mu\nu}$, as earlier, denotes the 4×4 diagonal matrix with diagonal entries $(-1, 1, 1, 1)$, associated with the Minkowski metric. We associate with $\mathcal{L}[\phi]$ the so-called *action integral*:

$$S[\phi] = \int_{\mathbb{R}^{3+1}} \mathcal{L}[\phi].$$

Notice that both $\mathcal{L}[\phi]$ and $S[\phi]$ are invariant under translations and Lorentz transformations. In other words, if $T : \mathbb{R}^{1+3} \rightarrow \mathbb{R}^{1+3}$ is a function that does not change the metric and we define a new function by $\psi(t, x) = \phi(T(t, x))$, then $\mathcal{L}[\phi] = \mathcal{L}[\psi]$ and $S[\phi] = S[\psi]$.

We shall consider a function ϕ that minimizes the action integral. From this we wish to deduce that its derivative, in some appropriate sense, is zero, and hence to deduce other properties about ϕ . But ϕ is a function that lives in an infinite-dimensional space, so we cannot talk about derivatives in a completely straightforward way. To deal with this problem, we define a *compact variation* of ϕ to be a smooth one-parameter family of functions $\phi^{(s)} : \mathbb{R}^{1+3} \rightarrow \mathbb{R}$, defined for each s in some interval $(-\epsilon, \epsilon)$, such that $\phi^{(0)}(x) = \phi(x)$ for every $x \in \mathbb{R}^3$ and $\phi^{(s)}(x) = \phi(x)$ for every (s, x) outside some bounded subset of \mathbb{R}^{1+3} . This allows us to differentiate with respect to s .

Given such a variation, we denote the derivative $d\phi^{(s)}/ds|_{s=0}$ by $\dot{\phi}$.

Definition. A field ϕ is said to be *stationary* with respect to S if, for any compact variation $\phi^{(s)}$ of ϕ , we have

$$\left. \frac{d}{ds} S[\phi^{(s)}] \right|_{s=0} = 0.$$

The variational principle. *The variational principle, or principle of least action, states that an acceptable solution of a given physical system must be stationary with respect to the action integral associated with the Lagrangian of the system.*

The variational principle enables us to associate with the given Lagrangian a system of PDEs, obtained from the fact that ϕ is stationary, called the *Euler-Lagrange equations*. We illustrate this by showing that the nonlinear wave equation in \mathbb{R}^{1+3} , namely

$$\square\phi - V'(\phi) = 0, \quad (56)$$

is the Euler-Lagrange equation associated with the Lagrangian (55). Given a compact variation $\phi^{(s)}$ of ϕ , we set $S(s) = S[\phi^{(s)}]$. Integration by parts gives

$$\begin{aligned} \frac{d}{ds} S(s) \Big|_{s=0} &= \int_{\mathbb{R}^{3+1}} [-m^{\mu\nu} \partial_\mu \dot{\phi} \partial_\nu \phi - V'(\phi) \dot{\phi}] \\ &= \int_{\mathbb{R}^{3+1}} \dot{\phi} [\square\phi - V'(\phi)]. \end{aligned}$$

In view of the action principle and the arbitrariness of $\dot{\phi}$ we infer that ϕ must satisfy equation (56). Thus (56) is indeed the Euler-Lagrange equation associated with the Lagrangian $\mathcal{L}[\phi] = m^{\mu\nu} \partial_\mu \phi \partial_\nu \phi - V(\phi)$.

One can similarly show that the Maxwell equations of electromagnetism—along with their beautiful extensions to the Yang-Mills equations, wave maps, and the Einstein equations of general relativity—are also variational. That is, they too can be derived from a Lagrangian.

Remark. The variational principle asserts only that the acceptable solutions of a given system are stationary: in general, we have no reason to expect that the desired solutions minimize or maximize the action integral. Indeed, this fails to be the case for systems that have a time dependence, such as the Maxwell equations, Yang-Mills equations, wave maps, and Einstein equations.

However, there is a large class of variational problems, corresponding to time-independent physical systems or geometric problems, for which the desired solutions *do* turn out to be extremal. The simplest example is that of geodesics in a Riemannian manifold M , which are minimizers¹⁷ with respect to length. More precisely, the *length functional* takes a curve γ that passes through two fixed points of M and associates with it its length $L(\gamma)$, which plays the role of an action integral. In this case a geodesic is not just a stationary point for the functional but a minimum. We also saw earlier that, according to the Dirichlet principle, solutions to the Dirichlet problem (53) minimize the Dirichlet integral (52). Another example is provided

by the minimal-surface equation (7), the solutions of which are minimizers of the area integral.

The study of minimizers of various functionals, i.e., action integrals, is a venerable subject in mathematics that goes under the name of *calculus of variations* (see VARIATIONAL METHODS [III.96] for further discussion).

Associated with the variational principle is another fundamental principle. A *conservation law* for an evolution PDE is a law that says that some quantity, typically an integral quantity depending on the solution, must remain constant over time, for every solution of the equation.

Noether's principle. *To any continuous one-parameter group of symmetries of the Lagrangian there corresponds a conservation law for the associated Euler-Lagrange PDE.*

Examples of such conservation laws are the familiar laws of conservation of energy, conservation of momentum, and conservation of angular momentum, all of which have important physical meaning. (Here, the one-parameter group of symmetries is just translations in time.) For example, in the case of equation (56), the law of conservation of energy takes the form

$$E(t) = E(0), \quad (57)$$

where the quantity $E(t)$, which equals

$$\int_{\Sigma_t} \left(\frac{1}{2} (\partial_t \phi)^2 + \frac{1}{2} \sum_{i=1}^3 (\partial_i \phi)^2 + V(\phi) \right) dx, \quad (58)$$

is called the *total energy* at time t . (The notation Σ_t stands for the set of all points (t, x, y, z) as (x, y, z) ranges over \mathbb{R}^3 .) Observe that (57) provides an extremely important a priori estimate for solutions to (56) in the case when $V \geq 0$. Indeed, if the energy of the initial data at $t = 0$ is finite (that is, if $E(0) < \infty$), then

$$\int_{\Sigma_t} \left((\partial_t \phi)^2 + \sum_{i=1}^3 (\partial_i \phi)^2 \right) \leq E(0).$$

We say that the energy identity (57) is *coercive*, which means that it leads to an absolute bound on all solutions with finite initial energy.

4.2 The Issue of Criticality

For the most basic evolution equations of mathematical physics, there are typically no better a priori estimates known than those provided by the energy. Taking into account the scaling properties of the corresponding equations as well, one is led to the very important classification of our basic equations, mentioned

17. This is true, in general, only for sufficiently short geodesics, i.e., ones that pass through two points close to each other.

earlier, into *subcritical*, *critical*, and *supercritical* equations. To see how this is done, consider again the nonlinear scalar equation $\square\phi - V'(\phi) = 0$, and take $V(\phi)$ to be $(1/(p+1))|\phi|^{p+1}$. Recall that the energy integral is given by (58). If we assign to the spacetime variables the dimension of length, L , then the spacetime derivatives have dimension L^{-1} and therefore \square has the dimension of L^{-2} . To be able to balance the left- and right-hand sides of the equation $\square\phi = |\phi|^{p-1}\phi$, we need to assign a length scale to ϕ ; we find this to be $L^{2/(1-p)}$. Thus the energy integral,

$$E(t) = \int_{\mathbb{R}^d} (2^{-1}|\partial\phi|^2 + |\phi|^{p+1}) dx,$$

has the dimension L^c , $c = d - 2 + (4/(1-p))$, with d corresponding to the volume element $dx = dx^1 dx^2 \cdots dx^d$, which scales like L^d . We say that the equation is *subcritical* if $c < 0$, *critical* if $c = 0$, and *supercritical* if $c > 0$. Thus, for example, $\square\phi - \phi^5 = 0$ is critical in dimension $d = 3$. The same sort of dimensional analysis can be done for all our other basic equations. An evolutionary PDE is said to be *regular* if all smooth finite-energy initial conditions lead to global smooth solutions. It is conjectured that all subcritical equations are regular, but one expects supercritical equations to develop singularities. Critical equations are important borderline cases. The heuristic reason for this is that the nonlinearity tends to produce singularities while the coercive estimates prevent it. In subcritical equations the coercive estimates are stronger, while for supercritical equations it is the nonlinearity that is stronger. However, there may be other, more subtle a priori estimates that are not accounted for by our crude heuristic argument. Thus, some supercritical equations, such as the Navier-Stokes equations, may still be regular.

4.3 Other Equations

Many other familiar equations can be derived from the variational ones described above by the following procedures.

4.3.1 Symmetry Reductions

Sometimes a PDE is very hard to solve but becomes much easier if one places additional symmetry constraints on solutions. For example, if the PDE is rotation invariant and we look just for rotation-invariant solutions $u(t, x)$, then we can regard these solutions as functions of t and $r = |x|$, effectively reducing the

dimension of the problem. By this procedure of *symmetry reduction* one can then derive a new PDE that is much simpler than the original one. Another, somewhat more general, way of obtaining simpler equations is to look for solutions that satisfy some further property. For instance, one can assume that they are stationary (that is, that they do not depend on the time variable), spherically symmetric, *self-similar* (which means that $u(t, x)$ depends only on x/t^a), or *traveling waves* (which means that $u(t, x)$ depends only on $x - vt$ for some fixed velocity vector v). Typically, the equations obtained by such reductions have a variational structure themselves. In fact, the symmetry reduction can be applied directly to the original Lagrangian.

4.3.2 The Newtonian Approximation and Other Limits

We can derive a large class of new equations as *limits* of the basic ones described above by taking one or more characteristic speeds to infinity. The most important example is the *Newtonian limit*, which is obtained by letting the velocity of light go to infinity. As we have already mentioned, the Schrödinger equation can be derived in this way from the linear Klein-Gordon equation. Similarly, we can derive the Lagrangians for the equations of nonrelativistic elasticity, fluid dynamics, or magnetohydrodynamics. It is an interesting fact that the nonrelativistic equations tend to look more messy than the relativistic ones. The simple geometric structure of the original equations gets lost in the limit. The remarkable simplicity of the relativistic equations is a powerful example of the importance of relativity as a unifying principle.

Once we are in the familiar world of Newtonian physics we can perform other well-known limiting procedures. The famous INCOMPRESSIBLE EULER EQUATIONS [III.23] are obtained by taking the limit of the general nonrelativistic fluid equations as the speed of sound tends to infinity. Various other limits are obtained relative to other characteristic speeds of the system or in connection with specific boundary conditions, such as the boundary-layer approximation in fluids. For example, in the limit as all characteristic speeds tend to infinity, the equations of elasticity turn into the familiar equations of a rigid body in classical mechanics.

4.3.3 Phenomenological Assumptions

Even after taking various limits and making symmetry reductions, the equations may still remain intractable.

However, in various applications it makes sense to assume that certain quantities are sufficiently small to be neglected. This leads to simplified equations that could be called *phenomenological*¹⁸ in the sense that they are not derived from first principles.

Phenomenological equations are “toy equations” that are used to illustrate and isolate important physical phenomena in complicated systems. A typical way of generating interesting phenomenological equations is to try to write down the simplest model equation that still exhibits a particular feature of the original system. For instance, the self-focusing plane-wave effects of compressible fluids or elasticity can be illustrated by the simple-minded Burger equation $u_t + uu_x = 0$. Nonlinear dispersive phenomena, typical of fluids, can be illustrated by the famous Korteweg–de Vries equation $u_t + uu_x + u_{xxx} = 0$. The nonlinear Schrödinger equation (54) provides a good model problem for nonlinear dispersive effects in optics.

If it is well chosen, a model equation can lead to basic insights into the original equation itself. For this reason, simplified model problems are also essential in the day-to-day work of the rigorous researcher into PDEs, who tests ideas on carefully selected model problems. It is crucial to emphasize that good results concerning the basic physical equations are rare; a very large percentage of important rigorous work in PDEs deals with simplified equations selected, for technical reasons, to isolate and focus our attention on some specific difficulties present in the basic equations.

In the above discussion we have not mentioned diffusive equations¹⁹ such as the Navier–Stokes equations. These are in fact not variational, and therefore do not quite fit into the above description. Though they could be viewed as phenomenological equations, they can also be derived from basic microscopic laws such as those governing the Newtonian–mechanical interactions of a very large number of particles N . In principle,²⁰ the equations of continuum mechanics, such as the Navier–Stokes equations, could be derived by letting the number of particles $N \rightarrow \infty$.

Diffusive equations also turn out to be very useful in connection with geometric problems. Geometric flows

such as mean curvature, inverse mean curvature, harmonic maps, Gauss curvature, and Ricci flow are some of the best-known examples. Diffusive equations can often be interpreted as the gradient flow for an associated elliptic variational problem. They can be used to construct nontrivial stationary solutions to the corresponding stationary systems, in the limit as $t \rightarrow \infty$, or to produce foliations with remarkable properties, such as one that was used recently in the proof of a famous conjecture of Penrose. As we have already mentioned, this idea has recently found an extraordinary application in the work of Perelman, who has used Ricci flow to settle the three-dimensional Poincaré conjecture. One of his main new ideas was to interpret Ricci flow as a gradient flow.

4.4 Regularity or Breakdown

An additional source of unity for the subject of PDEs is the central role played by the problem of *regularity or breakdown* of solutions to the basic equations. It is intimately tied to the fundamental mathematical question of understanding what we actually mean by solutions and, from a physical point of view, to the issue of understanding the limits of validity of the corresponding physical theories. Thus, in the case of the Burger equation, for example, the problem of singularities can be tackled by extending our concept of solutions to accommodate *shock waves*, which are solutions that are discontinuous across certain curves in the (t, x) -space. In this case one can define a function space of generalized solutions in which the IVP has unique, global solutions. Though the situation for more realistic physical systems is far less clear and far from being satisfactorily solved, the generally held opinion is that shock-wave-type singularities can be accommodated without breaking the boundaries of the physical theory at hand. The situation for singularities in general relativity is radically different. The singularities one expects there are such that no continuation of solutions is possible without altering the physical theory itself. The prevailing opinion here is that only a gravitational quantum field theory could achieve this.

5 General Conclusions

What, then, is the modern theory of PDEs? As a first approximation, one could say that it is the pursuit of the following main goals.

18. I use this term here quite freely; it is typically used in a somewhat different context. Also, some of the equations that I call phenomenological below, e.g., dispersive equations, can be given formal asymptotic derivations.

19. That is, equations where some of the basic physical quantities, such as energy, are not conserved and may in fact decrease in time. These are typically of parabolic type.

20. To establish this rigorously remains a major challenge.

(i) *Understand the problem of evolution for the basic equations of mathematical physics.* The most pressing issue in this regard is to understand *when and how the local*²¹ (with respect to time) *smooth solutions of the basic equations develop singularities.* A simple-minded criterion for distinguishing between regular theories and those that may admit singular solutions is given by the distinction between subcritical and supercritical equations. As mentioned earlier, it is widely believed that *subcritical equations are regular and that supercritical equations are not.* Indeed, many subcritical equations have been proved to be regular even though we lack a general procedure for establishing regularity results of this kind. The situation with supercritical equations is far more subtle. To start with, an equation that we now call supercritical²² may in fact turn out to be critical, or even subcritical, upon the discovery of additional a priori estimates. Thus an important question concerning the issue of criticality, and consequently that of singular behavior, is: are there other, stronger, local a priori bounds that cannot be derived from Noether's principle? The discovery of such a bound would be a major event in both mathematics and physics.

Once we understand that the presence of singularities in our basic evolution equations is unavoidable, we have to face the question of whether they can somehow be accommodated by a more general concept of what a solution is or whether their structure is such that the equation itself, indeed the physical theory that it underlies, becomes meaningless. An acceptable concept of a generalized solution should, of course, preserve the deterministic nature of the equations: in other words, it should be uniquely determined from its Cauchy data.

Finally, once an acceptable concept of generalized solutions is found, we would like to use it to determine some important qualitative features, such as long-term asymptotic behavior. One can formulate a limitless number of such questions, the answers to which will vary from equation to equation.

(ii) *Understand in a rigorous mathematical fashion the range of validity of various approximations.* The equations obtained by various limiting procedures or phenomenological assumptions can of course be stud-

ied in their own right, as the examples that we have referred to above are. However, they present us with additional problems to do with the mechanics of how they are derived from equations that we regard as more fundamental. It is entirely possible, for example, that the dynamics of a derived system of equations leads to behavior *that is incompatible with the assumptions made in its derivation.* Alternatively, a particular simplifying assumption, such as spherical symmetry in general relativity or zero vorticity for compressible fluids, may turn out to be unstable at large scales and therefore not a reliable predictor of the general case. These and other similar situations lead to important dilemmas: should we persist in studying the approximate equations even when, in many cases, we face formidable mathematical difficulties (some which may turn out to be quite pathological and are perhaps related to the nature of the approximation), or should we abandon them in favor of the original system or a more suitable approximation? Whatever one may feel about this in any specific situation, it is clear that the problem of understanding, rigorously, the range of validity of various approximations is one of the fundamental goals in PDEs.

(iii) *Devise and analyze the right equation for studying the specific geometric or physical problem at hand.* This last goal is equally important even though it is necessarily vague. The enormously important role played by PDEs in various branches of mathematics is more evident than ever. One looks in awe at how equations such as the Laplace, heat, wave, Dirac, KdV, Maxwell, Yang-Mills, and Einstein equations, which were originally introduced in specific physical contexts, turned out to have very deep applications to seemingly unrelated problems in areas such as geometry, topology, algebra, and combinatorics. Other PDEs appear naturally in geometry when we look for embedded objects with optimal geometric shapes, such as solutions to isoperimetric problems, minimal surfaces, surfaces of least distortion or minimal curvature, or, more abstractly, connections, maps, or metrics with distinguished properties. They are variational in character, just like the main equations of mathematical physics. Other equations have been introduced with the goal of allowing one to deform a general object, such as a map, connection, or metric, to an optimal one. They usually arise in the form of geometric, parabolic flows. The most famous example of this is Ricci flow, first introduced by Richard Hamilton, who hoped to use it to deform

PUP: Tim thinks
this is fine. OK?

21. One of the important achievements of the past century of mathematics was the establishment of a general procedure that guarantees the existence and uniqueness of a local-in-time solution to broad classes of initial conditions and large classes of nonlinear equations, including all those we have already mentioned above.

22. What we call supercritical depends on the strongest a priori coercive estimate available.

Riemannian metrics into Einstein metrics. Similar ideas were used earlier to construct, for example, stationary harmonic maps with the help of a harmonic heat flow, and self-dual Yang–Mills connections with the help of a Yang–Mills flow. In addition to the successful use of Ricci flow to settle the Poincaré conjecture in three dimensions, another remarkable recent example of the usefulness of geometric flows is that of the inverse mean flow, first introduced by Geroch, to settle the so-called Riemannian version of the Penrose inequality.

Further Reading

- Evans, L. C. 1998. *Partial Differential Equations*. Graduate Studies in Mathematics, volume 19. Providence, RI: American Mathematical Society.
- John, F. 1991. *Partial Differential Equations*. New York: Springer.
- Wald, R. M. 1984. *General Relativity*. Chicago, IL: Chicago University Press.
- Brezis, H., and F. Browder. 1998. Partial differential equations in the 20th century. *Advances in Mathematics* 135: 76–144.
- Constantin, P. 2007. On the Euler equations of incompressible fluids. *Bulletin of the American Mathematical Society* 44:603–21.
- Klainerman, S. 2000. PDE as a unified subject. In **GAFA 2000*, Visions in Mathematics—Towards 2000* (special issue of *Geometric and Functional Analysis*), part 1, pp. 279–315.

IV.13 General Relativity and the Einstein Equations

Mihalis Dafermos

Einstein’s formulation of general relativity represents one of the great triumphs of modern physics and provides the currently accepted classical theory that unifies gravitation, inertia, and geometry. The *Einstein equations* are the mathematical embodiment of this theory.

The definitive form of the equations,

$$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = 8\pi T_{\mu\nu}, \quad (1)$$

was attained in November 1915; this was the final act of Einstein’s eight-year struggle to generalize his *principle of relativity* so as to encompass gravitation, which had been described in the earlier “Newtonian” theory by the *Poisson equation*

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} = 4\pi\mu \quad (2)$$

for the potential ϕ and mass density μ .

An obvious contrast between the Einstein equations (1) and the Poisson equation (2) is that the mysterious notation of the former makes it far less obvious what they even mean. This has given the subject of general relativity a reputation for difficulty and impenetrability. However, this reputation is to some extent unwarranted. Both (1) and (2) represent the culmination of revolutionary theories whose formulations presuppose a complicated conceptual framework. For better or for worse, however, the structure necessary to formulate Poisson’s equation has been incorporated into our traditional mathematical notation and school education. As a result, \mathbb{R}^3 , with its Cartesian coordinate system, and notions such as functions, partial derivatives, masses, forces, and so on, are familiar to people with a general mathematical background, while the conceptual structure of general relativity is much less so, both with respect to its basic physical notions and with respect to the mathematical objects that are needed to model them. However, once one comes to terms with these, the equations turn out to be more natural and, one might even dare say, simpler.

Thus, the first task of this article is to explain in more detail the conceptual structure of general relativity. Our aim will be to make it clear what the equations (1) actually denote, and, moreover, why they are in a certain sense the simplest equations one can write down, given the general framework of the theory. This in turn will require us to review *special relativity* and its implications for the structure of matter, which will bring us to the unified concept of *stress-energy-momentum*, described by a *tensorial object* T . Finally, we will join Einstein in his inspired leap to the notion of a general four-dimensional *Lorentzian manifold* (\mathcal{M}, g) that represents our space-time continuum. We shall see that equation (1) expresses a relationship between the tensor T and the *geometry* of g as expressed in its so-called *curvature*.

There is more to truly understanding a theory than merely knowing how to write down its governing equations. General relativity is associated with some of the most spectacular predictions of twentieth-century physics: *gravitational collapse*, *black holes*, *space-time singularities*, the *expansion of the universe*. These phenomena (which were completely unknown in 1915 and thus played no role in the formulation of the equations (1)) revealed themselves only when the conceptual issues surrounding the problem of global *dynamics* of solutions were understood. This took a surprisingly

long time, though the story is not as well-known as the heroic struggle to attain (1). The article will conclude with a very brief glimpse into the fascinating dynamics of the Einstein equations.

1 Special Relativity

1.1 Einstein, 1905

Einstein's 1905 formulation of special relativity stipulated that all fundamental laws of physics should be invariant under *Lorentz transformations* of the *frame of reference* defined by x , y , z , and t . A Lorentz transformation is any composition of translations, rotations, and the *Lorentz boost*, which is given by the formulas

$$\left. \begin{aligned} \tilde{x} &= \frac{x - vt}{\sqrt{1 - v^2/c^2}}, & \tilde{y} &= y, \\ \tilde{t} &= \frac{t - vx/c^2}{\sqrt{1 - v^2/c^2}}, & \tilde{z} &= z, \end{aligned} \right\} \quad (3)$$

where c is a certain constant and $|v| < c$. Thus, Einstein's stipulation was that if one changes coordinates by means of a Lorentz transformation, then the form of all fundamental equations will remain the same. This set of transformations had already been identified in the context of the study of the vacuum *Maxwell equations* for the electric field \mathbf{E} and magnetic field \mathbf{B} :

$$\left. \begin{aligned} \nabla \cdot \mathbf{E} &= 0, & \nabla \cdot \mathbf{B} &= 0, \\ c^{-1} \partial_t \mathbf{B} + \nabla \times \mathbf{E} &= 0, & c^{-1} \partial_t \mathbf{E} - \nabla \times \mathbf{B} &= 0. \end{aligned} \right\} \quad (4)$$

Indeed, the Lorentz transformations are precisely the transformations that keep the form of the above equations invariant if we also transform \mathbf{E} and \mathbf{B} appropriately. Their significance was emphasized by POINCARÉ [VI.61]. However, it was Einstein's profound insight to elevate this invariance to the status of fundamental physical principle, despite its incompatibility with what we now usually call *Galilean relativity*, which corresponds to taking $c \rightarrow \infty$ in (3). A surprising consequence of Lorentz invariance is that the notion of simultaneity is not absolute but depends on the observer: given two distinct events that occur at (t, x, y, z) and (t, x', y', z') , it is easy to find a Lorentz transformation such that the transformed events no longer have the same t -coordinate.

It follows from a celebrated result in partial differential equations known as the *strong Huygens principle*, applied to (4), that electromagnetic disturbances in vacuum propagate with speed c , which we thus identify as the speed of light. In view of Lorentz invariance, this statement is independent of the frame! A further

postulate of the principle of relativity is that physical theories should not allow massive particles to move at speeds (as measured in any frame) greater than or equal to c .

1.2 Minkowski, 1908

Einstein's understanding of special relativity was "algebraic." It was MINKOWSKI [VI.64] who first understood its underlying geometric structure, namely, that the content of the principle was contained in the *metric element*

$$-c^2 dt^2 + dx^2 + dy^2 + dz^2 \quad (5)$$

defined on \mathbb{R}^4 with coordinates (t, x, y, z) . We call \mathbb{R}^4 endowed with the metric (5) *Minkowski space-time* and denote it \mathbb{R}^{3+1} . Points of \mathbb{R}^{3+1} are referred to as *events*. The expression (5) is classical notation for the *inner product* defined on tangent vectors $\mathbf{v} = (c^{-1}v^0, v^1, v^2, v^3)$, $\mathbf{w} = (c^{-1}w^0, w^1, w^2, w^3)$ on \mathbb{R}^4 by

$$\langle \mathbf{v}, \mathbf{w} \rangle = -v^0 w^0 + v^1 w^1 + v^2 w^2 + v^3 w^3. \quad (6)$$

The Lorentz transformations constitute precisely the *symmetry group* of the geometry defined by (5). Einstein's principle of relativity could now be understood as the principle that the fundamental equations of physics must refer to space-time only through geometric quantities: that is, quantities that can be defined purely in terms of the metric. For example, from this point of view the reason that the notion of absolute simultaneity is not allowed is that it depends on a privileged hyperplane through any given point of \mathbb{R}^{3+1} . But there are Lorentz transformations that preserve the metric and send this hyperplane to another one through the given point, so nothing in the metric can pick out one particular hyperplane. Note that if a physical theory makes use of geometric quantities only, then it is automatically invariant under Lorentz transformations: this observation renders many complicated calculations unnecessary.

Let us explore this geometric point of view further. Note that nonzero vectors \mathbf{v} are naturally classified by the inner product $\langle \cdot, \cdot \rangle$ into three types, called *timelike*, *null*, and *spacelike*, according to whether $\langle \mathbf{v}, \mathbf{v} \rangle < 0$, $\langle \mathbf{v}, \mathbf{v} \rangle = 0$, or $\langle \mathbf{v}, \mathbf{v} \rangle > 0$, respectively. Idealized point particles traverse curves γ through space-time; these are called the *world lines* of the corresponding particles. The postulate (referred to earlier) that speed in any frame of reference is bounded by the speed of light c can now be formulated as the following statement: *if γ is the world line of a particle, then the vector $d\mathbf{y}/ds$*

PUP: author insisted that this was how people in the field say this, although Tim and I also preferred the version with 'a' in it.

must be timelike. (Null lines correspond to light rays in the geometric optics limit of (4).) This statement is independent of the parameter s of \mathbf{y} , but for world lines we shall always assume that $dt/ds > 0$. To phrase this more geometrically, $\langle d\mathbf{y}/ds, (c^{-1}, 0, 0, 0) \rangle < 0$, which we interpret as the statement that \mathbf{y} is *future-directed*.

We can now define the “length” of the world line of a particle by

$$\begin{aligned} L(\mathbf{y}) &= \int_{s_1}^{s_2} \sqrt{-\langle \dot{\mathbf{y}}, \dot{\mathbf{y}} \rangle} ds \\ &= \int_{s_1}^{s_2} \sqrt{c^2 \left(\frac{dt}{ds} \right)^2 - \left(\frac{dx}{ds} \right)^2 - \left(\frac{dy}{ds} \right)^2 - \left(\frac{dz}{ds} \right)^2} ds. \end{aligned} \quad (7)$$

Classically, the above expression would have been written simply as

$$L(\mathbf{y}) = \int_{\mathbf{y}} \sqrt{-c^2 dt^2 + dx^2 + dy^2 + dz^2},$$

which explains the notation (5). We refer to the quantity $c^{-1}L(\mathbf{y})$ as *proper time*. This is the time that is relevant in local physical processes; in particular, if *you* are the particle traversing the world line \mathbf{y} , then $c^{-1}L(\mathbf{y})$ is the time that you will *feel*.

The metric (5) contains three-dimensional Euclidean geometry

$$dx^2 + dy^2 + dz^2,$$

restricted to $t = 0$, say. More interestingly, it also contains *non-Euclidean geometry*

$$\left(1 - \frac{x}{r}\right) dx^2 + \left(1 - \frac{y}{r}\right) dy^2 + \left(1 - \frac{z}{r}\right) dz^2$$

when it is restricted to the hypersurface $t = c^{-1}r = c^{-1}\sqrt{x^2 + y^2 + z^2}$. It is hard to overestimate how revolutionary the notion was that the time of physical processes (including our very sensations) and the length of measuring rods are two interdependent aspects of a geometric structure that naturally lives on a four-dimensional space-time continuum. Indeed, even Einstein initially rejected Minkowski space-time, preferring to retain the independent reality of a definite “space,” albeit a space with a relative notion of simultaneity. Only as a result of his search for general relativity did he realize that this view is fundamentally untenable. We shall return to this in section 3.

2 Relativistic Dynamics and the Unification of Energy, Momentum, and Stress

Besides the space-time concept and its geometrization, the principle of relativity led to a profound

rearrangement and unification of the fundamental concepts of dynamics: mass, energy, and momentum. Einstein’s celebrated relation between mass and energy in the rest frame,

$$E_0 = mc^2, \quad (8)$$

is the best-known expression of one aspect of this unification. This relation arises naturally when one attempts to generalize Newton’s second law $m(d\mathbf{v}/dt) = \mathbf{f}$ to a relation between 4-vectors in Minkowski space.

General relativity has to be formulated in terms of *fields* rather than particles. As a first step toward understanding it, let us look at continuous media. Now, instead of particles we consider *matter fields*; the unification of dynamical concepts encompasses what is known as *stress*, and its complete expression is embodied by the so-called *stress-energy-momentum tensor* \mathbf{T} . This tensor is fundamental to general relativity, so we have no choice but to familiarize ourselves with it. It will be the key to the form of the Einstein equations (1) as well as to the object on their right-hand side.

For each point $\mathbf{q} \in \mathbb{R}^{3+1}$, the stress-energy-momentum tensor field \mathbf{T} gives us a map

$$\mathbf{T} : \mathbb{R}_{\mathbf{q}}^4 \times \mathbb{R}_{\mathbf{q}}^4 \rightarrow \mathbb{R} \quad (9)$$

defined by the formula

$$\mathbf{T}(\mathbf{w}, \tilde{\mathbf{w}}) = \sum_{\alpha, \beta=0}^3 T_{\alpha\beta} w^\alpha \tilde{w}^\beta.$$

Here, $T_{\alpha\beta} = T_{\beta\alpha}$ for each α and β . By $\mathbb{R}_{\mathbf{q}}^4$ we mean the space of vectors *at* \mathbf{q} . (In Minkowski coordinates, we often identify \mathbb{R}^4 with $\mathbb{R}_{\mathbf{q}}^4$, but it will be important to distinguish between the two when considering arbitrary coordinates in section 3.2.) Bilinear maps of the form (9) are known as *covariant 2-tensors*.

If the only matter present is described by what is known as a *perfect fluid*, then the components of \mathbf{T} are given by

$$\begin{aligned} T_{00} &= (\rho + p)u^0u^0 - p, & T_{0i} &= (\rho + p)u^i u^0, \\ T_{ij} &= (\rho + p)u^i u^j + p\delta^{ij}, \end{aligned}$$

where \mathbf{u} is the 4-velocity, a timelike vector normalized such that $\langle \mathbf{u}, \mathbf{u} \rangle = -c^2$, ρ is the *mass-energy*, p is the *pressure*, and where $\delta_{ij} = 1$ if $i = j$, 0 if $i \neq j$, and i and j range over 1, 2, 3. Greek indices will range over 0, 1, 2, 3. We identify T_{00} with *energy*, T_{0i} with *momentum*, and T_{ij} with *stress*. These notions are clearly frame-dependent. Finally, observe that $\mathbf{T}(\mathbf{u}, \mathbf{u}) = \rho c^2$. This is the field-theoretic version of the famous equation (8).

In general, \mathbf{T} is derived from the totality of all the matter fields by constitutive functions that depend

on the nature of the matter fields and their interactions. We need not worry here about such things. But, regardless of the nature of the matter fields involved, we always postulate that the following equations are satisfied:

$$-\partial_0 T_{0\alpha} + \sum_{i=1}^3 \partial_i T_{i\alpha} = 0.$$

Defining $\nabla^0 = -\partial_0$, $\nabla^i = \partial_i$, and introducing the *Einstein summation convention*, under which summation is implicit when an index appears both upstairs and downstairs, we may rewrite this as

$$\nabla^\mu T_{\mu\nu} = 0. \quad (10)$$

These equations are Lorentz invariant.

The above relations embody the *conservation of stress-energy-momentum* at a differential level. Integrating (10) between homologous hypersurfaces and applying the Minkowski-space version of the divergence theorem, one obtains global balance laws. If one assumes that $T_{\alpha\beta}$ is compactly supported, then, integrating between $t = t_1$ and $t = t_2$, one obtains

$$\int_{t=t_2} T_{0\alpha} dx^1 dx^2 dx^3 = \int_{t=t_1} T_{0\alpha} dx^1 dx^2 dx^3. \quad (11)$$

With respect to the chosen Lorentz frame, the zeroth component of the above equation represents the *conservation of total energy*, while the remaining components represent *conservation of total momentum*.

In the case of a perfect fluid, if we close the system (10) by adjoining a conservation law for particle number

$$\nabla^\alpha (n u_\alpha) = 0$$

and postulate constitutive relations between ρ , p , particle number density n , and entropy per particle s , compatible with the laws of thermodynamics, then we arrive at the so-called *relativistic Euler equations*.

3 From Special to General Relativity

With the elements of special relativity at hand, together with their deep implications for the nature of energy, momentum, and stress, we can now pass to the formulation of general relativity.

3.1 The Equivalence Principle

Einstein understood as early as 1907 that the most profound aspect of the gravitational force could not be described within the relativity principle as he had formulated it in 1905. This aspect is what he called *the equivalence principle*.

The easiest setting in which to understand this principle is that of the “test particle” with velocity $\mathbf{v}(t)$ in a fixed gravitational field ϕ . In this case, we have that the classical *gravitational force* is given by $\mathbf{f} = -m\nabla\phi$, and we may rewrite Newton’s second law $m(d\mathbf{v}/dt) = \mathbf{f}$ as

$$\frac{d\mathbf{v}}{dt} = -\nabla\phi. \quad (12)$$

Notice that the mass m has dropped out! Thus, the gravitational field accelerates all objects at a given position in the same way. This explains the fact, recorded already in late antiquity by Ioannes Philoponus and popularized in Western Europe by Galileo, that the time it takes objects to fall from a given height is independent of their weight.

It was Einstein who first interpreted this property as a sort of covariance with respect to transformations to *noninertial*, that is to say *accelerated*, frames. For instance, in the case of a constant gravitational field, which corresponds to the case $\phi(z) = fz$, we can pass to the accelerated frame

$$\tilde{z} = z + \frac{1}{2}ft^2$$

and write (12) as

$$\frac{d\mathbf{v}}{dt} = 0. \quad (13)$$

Similarly, one can reverse the argument to “simulate” a gravitational field when none is present by expressing (13) in an accelerated frame.

3.2 Vectors, Tensors, and Equations in General Coordinates

Exactly what the equivalence principle means in general is somewhat obscure and has been the subject of debate ever since Einstein introduced it. Nevertheless, the above considerations suggest that, even in the absence of gravity, it would be useful to know how various objects and equations appear when expressed in arbitrary coordinate systems. That is to say, let us change from our Minkowski coordinates x^0, x^1, x^2, x^3 to the most general coordinate system, which we shall write as $\tilde{x}^{\tilde{\mu}} = \tilde{x}^{\tilde{\mu}}(x^0, x^1, x^2, x^3)$, where $\tilde{\mu}$ ranges over 0, 1, 2, 3.

Expressing scalar functions in arbitrary coordinates poses no problem. But what about vector fields? If \mathbf{v} is a vector field expressed in Minkowski coordinates as (v^0, v^1, v^2, v^3) , how do we express \mathbf{v} in our new coordinates $\tilde{x}^{\tilde{\mu}}$?

One has to think a bit about what a vector field actually is. The correct point of view is to consider a vector field \mathbf{v} as a first-order differential operator defined

(using Einstein's summation convention) by $\mathbf{v}(f) = v^\mu \partial_\mu f$. So we seek v^μ such that $\mathbf{v}(f) = v^\mu \partial_{\bar{\mu}} f$ for all functions f . The chain rule then gives us our answer:

$$v^{\bar{\mu}} = \frac{\partial \bar{x}^{\bar{\mu}}}{\partial x^\nu} v^\nu. \quad (14)$$

What about tensors, such as the stress-energy-momentum tensor \mathbf{T} ? In view of the definition (9), we seek $T_{\bar{\mu}\bar{\nu}}$ such that

$$\mathbf{T}(\mathbf{u}, \mathbf{v}) = T_{\bar{\mu}\bar{\nu}} u^{\bar{\mu}} v^{\bar{\nu}}, \quad (15)$$

where the numbers $u^{\bar{\mu}}$ are the components of \mathbf{u} with respect to the coordinates $\bar{x}^{\bar{\mu}}$ as we have just calculated them above. (Note that these components depend on the point \mathbf{q} . This is why it is now essential to distinguish \mathbb{R}_q^4 from \mathbb{R}^4 .) Again, the chain rule gives us the answer:

$$T_{\bar{\mu}\bar{\nu}} = T_{\mu\nu} \frac{\partial x^\nu}{\partial \bar{x}^{\bar{\mu}}} \frac{\partial x^\mu}{\partial \bar{x}^{\bar{\nu}}}.$$

Classically, we write

$$\mathbf{T} = T_{\bar{\mu}\bar{\nu}} d\bar{x}^{\bar{\mu}} d\bar{x}^{\bar{\nu}} = T_{\mu\nu} dx^\mu dx^\nu.$$

One can interpret the above as a shorthand notation for (15), but it also tells us how to compute $T_{\bar{\mu}\bar{\nu}}$ from $T_{\mu\nu}$ by formally applying the chain rule to $d\bar{x}^{\bar{\mu}}$.

There is another covariant symmetric 2-tensor besides \mathbf{T} that is relevant here. This is the Minkowski metric itself. Indeed, the classical form of the Minkowski metric (5) corresponds to the representation

$$\eta_{\mu\nu} dx^\mu dx^\nu,$$

where the $\eta_{\mu\nu}$ for Minkowski coordinates x^μ are given by $\eta_{00} = -1$, $\eta_{0i} = 0$, $\eta_{ij} = 1$ if $i = j$, and $\eta_{ij} = 0$ if $i \neq j$. To avoid the cumbersome notation $\langle \cdot, \cdot \rangle$, let us refer to the Minkowski metric as $\boldsymbol{\eta}$. Following the above, we may express $\boldsymbol{\eta}$ in general coordinates $\bar{x}^{\bar{\mu}}$ by

$$\eta_{\bar{\mu}\bar{\nu}} d\bar{x}^{\bar{\mu}} d\bar{x}^{\bar{\nu}},$$

where $\eta_{\bar{\mu}\bar{\nu}}$ is computed by formal application of the chain rule.

It is clear that if one tries to transform an equation such as (10) into general coordinates, then the components of $\boldsymbol{\eta}$ and their derivatives will appear in the equations. Einstein (always thinking “algebraically”) was seeking laws of motion for both matter and the gravitational field that would have the same *form* in all coordinate systems. As he understood it, this meant that all objects that appear should transform as tensors and should be considered a priori “unknown.” He referred to this principle as “general covariance.” This suggests that $\boldsymbol{\eta}$ should be replaced by an *unknown* symmetric 2-tensor. Let us call this 2-tensor \mathbf{g} . One can of course try to write down an equation for the “unknown”

\mathbf{g} that forces it to be the “known” Minkowski metric $\boldsymbol{\eta}$. Thus, “general covariance” per se does not force one to abandon $\boldsymbol{\eta}$. But in view of the fact that \mathbf{g} and \mathbf{T} have the same number of components, it was a natural step to consider \mathbf{g} as the embodiment of the gravitational field and to try to look for an equation that related \mathbf{g} and \mathbf{T} directly. In this way, the framework of general relativity was born.

3.3 Lorentzian Geometry

The profound insight of replacing the fixed Minkowski $\boldsymbol{\eta}$ with a dynamic \mathbf{g} brought Einstein to what we now call *Lorentzian geometry*. Lorentzian geometry generalizes Minkowski geometry following the blueprint of RIEMANN [VI.49]. That is, we replace the Minkowski metric $\boldsymbol{\eta}$ by a general map

$$\mathbf{g} : \mathbb{R}_q^4 \times \mathbb{R}_q^4 \rightarrow \mathbb{R}.$$

In other words, we replace $\boldsymbol{\eta}$ by a symmetric covariant 2-tensor, which is expressed in arbitrary coordinates x^μ by

$$g_{\mu\nu} dx^\mu dx^\nu.$$

Moreover, we require that at each point \mathbf{q} the bilinear form $\mathbf{g}(\cdot, \cdot)$ can be diagonalized to the Minkowski form (6). Loosely speaking, a Lorentzian metric is one that “looks locally like the Minkowski metric,” just as a RIEMANNIAN METRIC [I.3 §6.10] looks locally like the Euclidean metric.

Just as with the Minkowski metric, the bilinear form \mathbf{g} permits us to classify nonzero vectors \mathbf{v}_q at a point \mathbf{q} as *timelike*, *null*, or *spacelike* and to define proper times of world lines $\gamma(s) = (x^0(s), x^1(s), x^2(s), x^3(s))$ by the formula (7), but with $\langle \dot{\gamma}, \dot{\gamma} \rangle$ replaced by $g_{\mu\nu} \dot{x}^\mu \dot{x}^\nu$. It is in this sense that we can speak of the *geometry* of \mathbf{g} .

In view of Minkowski's formulation of the special relativity principle as the statement that the equations of physics refer to space-time only through geometric quantities associated with the Minkowski metric, it is natural to look for a generalization of this principle, and indeed a suitable version immediately suggests itself. It is the principle that *the equations of physics refer to the space-time coordinates only via geometric quantities naturally associated with \mathbf{g}* .

The kinematic constraint on “test particles” as formulated geometrically for the Minkowski metric, namely that $d\mathbf{y}/ds$ should be timelike, makes sense for an arbitrary Lorentzian metric. But how does one formulate differential equations? For instance, how does one formulate an analogue of (10) that refers only to \mathbf{g} ?

It turned out that in the Riemannian case, a set of natural geometric concepts suitable for the task had already been developed in the nineteenth and early twentieth centuries by Riemann, Bianchi, Christoffel, Ricci, and Levi-Civita. These carry over directly to the Lorentzian case.

One begins by defining the so-called *Christoffel symbols* $\Gamma_{\mu\nu}^\lambda$ by

$$\Gamma_{\mu\nu}^\lambda = \frac{1}{2} g^{\lambda\rho} (\partial_\mu g_{\rho\nu} + \partial_\nu g_{\mu\rho} - \partial_\rho g_{\mu\nu}).$$

Here, the numbers $g^{\mu\nu}$ are the components of the “inverse metric” of g : that is, they are the unique solution to the equation $g^{\mu\nu} g_{\nu\lambda} = \delta_\lambda^\mu$, where, as usual, $\delta_\lambda^\mu = 1$ if $\lambda = \mu$ and 0 otherwise. (It turns out that $g^{\mu\nu}$ is very useful for the calculational gymnastics that are typical of tensor analysis when it exploits the Einstein summation convention.)

One can then define a differential operator ∇_μ called a *connection*, which acts on vector fields by

$$\nabla_\mu v^\nu = \partial_\mu v^\nu + \Gamma_{\mu\lambda}^\nu v^\lambda \quad (16)$$

and on covariant 2-tensors by

$$\nabla_\lambda T_{\mu\nu} = \partial_\lambda T_{\mu\nu} - \Gamma_{\lambda\mu}^\sigma T_{\sigma\nu} - \Gamma_{\lambda\nu}^\sigma T_{\mu\sigma}. \quad (17)$$

The left-hand sides of (16) and (17) define tensors that can be expressed in any coordinate system by a formal application of the chain rule.

With the help of this differential operator, one could now write the analogue of equations (10) for an arbitrary metric g as

$$\nabla^\mu T_{\mu\nu} = 0, \quad (18)$$

where $\nabla^\mu = g^{\mu\nu} \nabla_\nu$ refers to the connection associated with g .

If we consider a limit as the matter field becomes concentrated at a point, or rather as the stress-energy-momentum tensor $T_{\mu\nu}$ is nonzero only on a world line, then this curve will be a *geodesic* of g : that is, a curve that locally maximizes the proper time defined by g . These are the analogues of straight timelike lines in Minkowski space. In this limit, the motion of the matter does not depend on the nature of the stress-energy-momentum tensor, but only on the geometry of the metric that defines geodesics. Thus, all objects fall in the same way. These considerations give a concrete realization to the equivalence principle in general relativity.

Finally, it is important to remark that for a general metric g , the identity (18) *does not* imply global conservation laws (11) for “total energy” and “total momentum.” Such laws hold only if g has symmetries. The

fact that the fundamental conservation laws survive in general only at the infinitesimal level is an important insight into the nature of these principles in physics.

3.4 Curvature and the Einstein Equations

It remains, then, to give a set of equations for the metric g that relate it to T . In anticipation of a Newtonian limit, we expect these equations to be second order, and we expect them to implement “general covariance” in the simplest way possible: they should refer to no other structure but g itself and T .

Again, Riemannian geometry provides ready-made tensorial objects that are invariantly associated with g . One can define the *Riemann curvature tensor*

$$R_{\mu\nu\lambda\rho} dx^\mu dx^\nu dx^\lambda dx^\rho$$

with components given by

$$R_{\mu\nu\lambda\rho} = g_{\mu\sigma} (\partial_\rho \Gamma_{\nu\lambda}^\sigma - \partial_\lambda \Gamma_{\nu\rho}^\sigma + \Gamma_{\nu\lambda}^\tau \Gamma_{\tau\rho}^\sigma - \Gamma_{\nu\rho}^\tau \Gamma_{\tau\lambda}^\sigma).$$

One can also define the *Ricci curvature*

$$R_{\mu\nu} dx^\mu dx^\nu,$$

a covariant symmetric 2-tensor with components given by

$$R_{\mu\nu} = g^{\lambda\rho} R_{\mu\nu\lambda\rho},$$

and the *scalar curvature*

$$R = g^{\mu\nu} R_{\mu\nu}.$$

If g were the induced (Riemannian) metric on a 2-surface in \mathbb{R}^3 , then R would just be twice the *Gauss curvature* K . The above expressions should be thought of as complicated tensorial generalizations of Gauss curvature to several dimensions.

The final piece of the puzzle for the formulation of the Einstein equations (1) is provided by the following constraint that Einstein demanded: whatever the equation relating the metric and the stress-energy-momentum tensor of matter, (18) (the infinitesimal conservation of stress-energy-momentum) should hold *as a consequence*. Now, it turns out that for *any* metric g , the so-called *Bianchi identities* imply that

$$\nabla^\mu (R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R) = 0. \quad (19)$$

It is thus natural to postulate a linear relation between $T_{\mu\nu}$ and the tensor $R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R$. The form

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = 8\pi G c^{-4} T_{\mu\nu} \quad (20)$$

is then uniquely determined by the requirement that it should give the correct Newtonian limit when one makes the identifications

$$g_{00} \sim 1 + 2\phi/c^2, \quad g_{0j} \sim 0, \quad g_{ij} \sim (1 - 2\phi/c^2) \delta_{ij}.$$

The form (1) corresponds to the usual units $G = c = 1$. Note that (1), when written out explicitly, is nonlinear in the metric components $g_{\mu\nu}$.

Einstein did not stop at the Newtonian limit. By considering geodesic motion in solutions of the linearized equations (20), Einstein was able to determine the correct value for the *anomalous precession of the perihelion of Mercury*, an effect that Newtonian theory was unable to explain. Since (20) had no adjustable parameters after determining the Newtonian limit, this was a genuine *test* of the theory. A few years later the gravitational “bending” of light was observed. This had been calculated theoretically in the context of the geometric optics approximation where light rays follow null geodesics in a fixed space-time background. Post-Newtonian predictions of (1) have now been verified by various solar system tests, confirming general relativity in this regime to a high degree of accuracy.

One special case of (20) is when we postulate that $T_{\mu\nu} = 0$. The equations then simplify to

$$R_{\mu\nu} = 0. \quad (21)$$

These are known as the *vacuum equations*. The Minkowski metric (5) is a particular solution (but not the only one!).

The vacuum equations can be derived formally as the EULER-LAGRANGE EQUATIONS [III.96] corresponding to the so-called *Hilbert Lagrangian*:

$$\mathcal{L}(g) = \int R \sqrt{-g} dx^0 dx^1 dx^2 dx^3.$$

(The expression $\sqrt{-g} dx^0 dx^1 dx^2 dx^3$ denotes the natural *volume form* associated with g .) HILBERT [VI.63], who was following closely Einstein’s struggle to formulate a theory of gravity with a dynamic metric g , arrived at his Lagrangian (actually a more general version of the above yielding the coupled Einstein-Maxwell system) very shortly before Einstein obtained the general equations (20).

Many of the most interesting phenomena that come from the equations (20) are already present in the vacuum case (21). This is somewhat ironic, because it was the forms of T and (10) that dictated (20). Note, in contrast, that in the Newtonian theory (2), the “vacuum” equations $\mu = 0$ and standard boundary conditions at infinity imply $\phi = 0$. Thus, the Newtonian theory of the vacuum is trivial.

The part of the curvature tensor $R_{\mu\nu\lambda\rho}$ that is not forced to vanish from (21) is known as the *Weyl curvature*. This curvature measures the “tidal” distortion of families of geodesics. Thus, the “local strength” of

gravitational fields in vacuum regions is related in the Newtonian limit to the tidal forces on macroscopic test matter, not the norm of the gravitational force.

3.5 The Manifold Concept

We have been able to get this far without really addressing the question of *where the metric g is defined*. In passing from the Minkowski metric to a general g , Einstein did not originally have in mind replacing the domain \mathbb{R}^4 . But it is clear in the Riemannian case from the theory of surfaces that the natural object for a metric to live on is not necessarily \mathbb{R}^2 but a general surface. For instance, the metric $d\theta^2 + \sin^2\theta d\phi^2$ naturally lives on the sphere S^2 . In saying this, we are to understand that one requires several coordinate systems of the type (θ, ϕ) to cover all of S^2 . The n -dimensional generalization of the object where Riemannian or Lorentzian metrics naturally live is a MANIFOLD [I.3 §6.9]. Manifolds are the structures obtained by consistently smoothly pasting together local coordinate systems.

Thus, general relativity allows the space-time continuum not to be \mathbb{R}^4 but instead to be a general manifold \mathcal{M} , which may very well be topologically inequivalent to \mathbb{R}^4 , just as S^2 is inequivalent to \mathbb{R}^2 . We call the pair (\mathcal{M}, g) a *Lorentzian manifold*. Properly put, the unknown in the Einstein equations is not just g but the pair (\mathcal{M}, g) .

It is interesting that this fundamental fact, namely that the topology of space-time is not a priori determined by the equations, arises almost as an afterthought. Moreover, it was a thought that took many years to be clarified.

3.6 Waves, Gauges, and Hyperbolicity

When written out explicitly in arbitrary coordinates (try it!), the Einstein equations do not appear to be of any usual type, such as elliptic (like THE POISSON EQUATION [IV.12 §1]), parabolic (like THE HEAT EQUATION [I.3 §5.4]), or hyperbolic (like THE WAVE EQUATION [I.3 §5.4]; see [IV.12 §2.5] for more about these different classes of PDEs). This is related to the fact that, given a solution, one can form a “new” solution by composing the old solution with a coordinate transformation. We can do this for new coordinate systems whose coordinate transformations differ from the identity only in a ball. This fact, known as the *hole argument*, confused Einstein and his mathematical collaborator Marcel Grossmann, who were thinking algebraically in

terms of the form of the equations in coordinates, and temporarily led them to reject “general covariance.” The resulting backtracking delayed the final correct formulation of (1) by about two years. The geometric interpretation of the theory immediately suggests the resolution to the dilemma: such solutions are to be considered “the same” because they are the same from the point of view of all geometric measurements. In modern language, a solution to the Einstein vacuum equations (say) is an EQUIVALENCE CLASS [I.2 §2.3] of space-times (\mathcal{M}, g) , where two space-times are equivalent if there exists a diffeomorphism ϕ between them such that in any open set the metric has the same coordinate form when one identifies local coordinates by ϕ .

It turns out that once these conceptual issues are overcome, the Einstein equations can be viewed as hyperbolic. The easiest way to do this is to impose a *gauge*: that is to say, a certain restriction on the coordinate system. Specifically, one requires the coordinate functions x^α to satisfy the wave equation $\square_g x^\alpha = 0$, where the *d'Alembertian* operator is defined by the formula

$$\square_g = \frac{1}{\sqrt{-g}} \partial_\mu (\sqrt{-g} g^{\mu\nu} \partial_\nu).$$

Such coordinates always exist locally and they are traditionally called *harmonic coordinates*, although the term *wave coordinates* would perhaps be more appropriate. The Einstein equation can then be written as a system

$$\square_g g_{\mu\nu} = N_{\mu\nu}(\{g_{\alpha\beta}\}, \{\partial_\gamma g_{\alpha\beta}\}),$$

where $N_{\mu\nu}$ is a nonlinear expression that is quadratic in the $\partial_\gamma g_{\alpha\beta}$. In view of the Lorentzian signature of the metric, the above system constitutes what is known as a *second-order nonlinear (but quasilinear) hyperbolic system*.

At this point, it is instructive to make a comparison with the Maxwell equations. Suppose we are given an electric field E and a magnetic field B defined on Minkowski space. A *4-potential* is a vector field A such that $E_i = -\nabla_i A_0 - c^{-1} \partial_t A_i$, and $B_i = \sum_{j,k=1}^3 \epsilon_{ijk} \partial_j A_k$. (Here $\epsilon_{123} = 1$, and ϵ_{ijk} is totally antisymmetric, i.e., it transforms to its negative under permutation of any two indices.) If one wishes to view A as the fundamental physical object, then one notices that if A is replaced by the field \tilde{A} , defined by the formula

$$\tilde{A} = A + (-c^{-1} \partial_t \psi, \partial_1 \psi, \partial_2 \psi, \partial_3 \psi),$$

where ψ is an *arbitrary* function, then \tilde{A} is also a 4-potential for E and B . One can expect a determined equation for A only if one imposes further conditions

on it: that is, if one “fixes the gauge.” (The terminology “gauge” is originally due to WEYL [VI.80].) In the so-called *Lorentz gauge*

$$\nabla^\mu A_\mu = 0,$$

the Maxwell equations can be written

$$\square A_\mu = -c^{-2} \partial_t^2 A_\mu + \sum_i \partial_{x^i}^2 A_\mu = 0,$$

from which the wave properties are completely manifest. The gauge-symmetric point of view lived on to later twentieth century glory: the *Yang-Mills equations*, which are a nonlinear generalization of the Maxwell equations with a similar gauge symmetry, are the central part of the so-called *standard model* for particle physics.

The hyperbolicity property of the Einstein equations has two important repercussions. The first is that there should exist *gravitational waves*. This was noted by Einstein at least as early as 1918, essentially as a result of a linearized version of the considerations in the above discussion. The second is that there is a WELL-POSED INITIAL-VALUE PROBLEM [IV.12 §2.4] for the Einstein equations (1) with the domain-of-dependence property, when these are coupled with appropriate matter equations. In particular, this is true in the vacuum case (21). The proper conceptual framework to formulate the latter problem took a long time to get right, and was only completely understood through work of Choquet-Bruhat and Geroch in the 1950s and 1960s, based on the fundamental concept of *global hyperbolicity* due to Leray. Well-posedness means that one could associate a unique solution (in the vacuum case, a Lorentzian 4-manifold (\mathcal{M}, g) satisfying (21)) with a suitable notion of initial data. Of course, “initial data” does not mean “data at time $t = 0$,” since the concept of $t = 0$ is not geometric. Instead, the data take the form of some Riemannian 3-manifold (Σ, \bar{g}) with a symmetric covariant 2-tensor K . The triple (Σ, \bar{g}, K) has to satisfy the so-called *Einstein constraint equations*. But with this notion, the fundamental problem of general relativity, despite its revolutionary conceptual structure, is thoroughly classical: to determine the relation of the solution to initial data, that is to say, to determine the future from knowledge of the “present.” This is the problem of *dynamics*.

4 The Dynamics of General Relativity

In this final section we give a taste of our current mathematical understanding of the dynamics of the Einstein equations.

4.1 Stability of Minkowski Space and the Nonlinearity of Gravitational Radiation

In any physical theory in which one can formulate the problem of dynamics, the most basic question is the stability of the trivial solution. In other words, if we make a small change to the “initial conditions,” will the resulting change to the solution be small as well? In the case of general relativity, this is the question of stability of the Minkowski space-time \mathbb{R}^{3+1} . This fundamental result was proven for the vacuum equations (21) in 1993 by Christodoulou and Klainerman.

The proof of the stability of Minkowski space made it possible to formulate the *laws of gravitational radiation* rigorously. Gravitational radiation is yet to be observed directly, but it has been inferred, originally by Hulse and Taylor, from the energy loss of a binary system. This work gave them the only Nobel prize (1993) directly associated with the Einstein equations! The blueprint for the mathematical formulation of the radiation problem is based on work of Bondi and later Penrose. One associates with the space-time (\mathcal{M}, g) an ideal boundary “at infinity,” known as *null infinity* and denoted \mathcal{I}^+ . Physically, the points of \mathcal{I}^+ correspond to observers who are far away from the isolated self-gravitating system but who are receiving its signals. Gravitational radiation can be identified with certain tensors defined on \mathcal{I}^+ from rescaled boundary limits of various geometric quantities. As Christodoulou was to discover, the laws of gravitational radiation are themselves nonlinear, and the nonlinearity is potentially relevant for observation.

4.2 Black Holes

Perhaps no prediction of general relativity is better known today than that of black holes.

The story of black holes begins with the so-called *Schwarzschild* metric:

$$- \left(1 - \frac{2m}{r}\right) dt^2 + \left(1 - \frac{2m}{r}\right)^{-1} dr^2 + r^2 (d\theta^2 + \sin^2 \theta d\phi^2). \quad (22)$$

The parameter m here is a positive constant. This is a solution of the vacuum Einstein equations (21) that was found in 1916. The original interpretation of (22) was that it modeled the gravitational field in a vacuum region outside a star. That is to say, (22) was considered only in some coordinate range $r > R_0$, for an $R_0 > 2m$, and the metric was matched at $r = R_0$ to a “static” interior metric satisfying the coupled Einstein-Euler sys-

tem in the coordinate range $r \leq R_0$. (This latter metric is again of the form (22), but with $m = m(r)$ such that $m \rightarrow 0$ as $r \rightarrow 0$.)

From the theoretical point of view, a natural problem poses itself. Suppose we do away with the star altogether and try to consider (22) for *all* values of r . What happens then to the metric (22) at $r = 2m$? In the (r, t) coordinates, the metric element appears to be singular. But this turns out to be an illusion! By a simple change of coordinates, one can easily extend the metric regularly as a solution of (21) beyond $r = 2m$. That is, there exists a manifold \mathcal{M} that contains both a region $r > 2m$ and a region $0 < r < 2m$, separated by a regular (null) hypersurface \mathcal{H}^+ . The metric element (22) is valid everywhere except on \mathcal{H}^+ , where it must be rewritten in regular coordinates.

It turns out that the hypersurface \mathcal{H}^+ can be characterized by an exceptional global property: it defines the boundary of the region of space-time that can send signals to null infinity \mathcal{I}^+ , or, in the physical interpretation, to distant observers. In general, the set of points that *cannot* send signals to null infinity \mathcal{I}^+ is known as the *black hole* region of space-time. Thus, the region $0 < r < 2m$ is the black hole region of \mathcal{M} , and \mathcal{H}^+ is known as the *event horizon*.

These issues took a long time to be sorted out, partly because the language of global Lorentzian geometry was developed long after the original formulation of the Einstein equations. The global geometry of the extended space-time \mathcal{M} was clarified by Synge in around 1950 and finally by Kruskal in 1960. The name “black hole” is due to the imaginative physicist John Wheeler. From their beginnings as a theoretical curiosity, black holes have become part of the accepted astrophysical explanation for a wide variety of phenomena, and in particular are thought to represent the end-state for the gravitational collapse of many stars.

4.3 Space-Time Singularities

A second natural problem poses itself in relation to the Schwarzschild metric (22), now considered in the region $r < 2m$ of the extended space-time \mathcal{M} : *what happens at $r = 0$?*

A computation reveals that as $r \rightarrow 0$, the Kretschmann scalar $R_{\mu\nu\lambda\rho} R^{\mu\nu\lambda\rho}$ blows up. Since this expression is a geometric invariant, it follows that, unlike the situation at $r = 2m$, the space-time is *not* regularly extendable beyond 0. Moreover, timelike geodesics (freely falling observers in the test particle approximation) entering

the black hole region reach $r = 0$ in finite proper time, so they are “incomplete” in the sense that they cannot be continued indefinitely. They thus “observe” the breakdown of the geometry of the space-time metric. Moreover, macroscopic observers approaching $r = 0$ are torn apart by the gravitational “tidal forces.”

In the early years of the subject, it was thought that this seemingly pathological behavior was connected to the high degree of symmetry of the Schwarzschild metric and that “generic” solutions would not exhibit such phenomena. That this is not the case was shown by Penrose’s celebrated *incompleteness theorem* of 1965. This states that solutions to the initial-value problem for the Einstein equations coupled to appropriate matter will *always* contain such incomplete timelike or null geodesics if the initial data hypersurface is noncompact and contains what is known as a closed trapped surface. The Schwarzschild case may appear to suggest that such incomplete geodesics are associated with the curvature blowing up. However, the situation can in fact be very different, as is apparent in the celebrated *Kerr* solutions, a remarkable two-parameter family of solutions to the vacuum equations (21), discovered only in 1963, which are rotating versions of (22). In the Kerr solutions, incomplete timelike geodesics meet a so-called *Cauchy horizon*, a smooth boundary of the region of space-time that is uniquely determined by initial data.

The theorem of Penrose gives rise to two important conjectures. The first, known as *weak cosmic censorship*, says roughly that for generic physically plausible initial data for suitable Einstein-matter systems, geodesic incompleteness, if it occurs, is always confined to black hole regions. The second, *strong cosmic censorship*, says roughly that for generic admissible initial data, incompleteness of the solution is always associated with a local obstruction to extendability, such as the blow-up of curvature. The latter conjecture would ensure that the unique solution of the initial-value problem is the only classical space-time that can arise from the data. That is to say, it would imply that classical determinism holds for the Einstein equations.

Both conjectures are false if we drop the assumption that the initial data are generic, and this is one reason for their difficulty. Indeed, Christodoulou has constructed spherically symmetric solutions of the coupled Einstein-scalar field system (arising from regular initial data) that are geodesically incomplete but do not contain black hole regions. Such space-times are said to contain *naked singularities*.

Naked singularities are easy to construct if one does not require that they arise from the collapse of regular initial data. An example is the Schwarzschild metric (22) for $m < 0$. This metric, however, does not admit a complete asymptotically flat Cauchy hypersurface. This fact is related to the celebrated *positive energy theorem* of Schoen and Yau.

4.4 Cosmology

The space-times (\mathcal{M}, g) discussed previously are all idealized representations of isolated systems. The “rest of the universe” is excised and replaced by an “asymptotically flat end”; far-away observers are placed at an ideal boundary “at infinity.” But what if we are more ambitious and consider our space-time (\mathcal{M}, g) as representing the whole universe? The study of this latter problem is known as *cosmology*.

Observations suggest that on very large scales the universe is approximately homogeneous and isotropic. This is sometimes known as the *Copernican principle*. Interestingly, one cannot solve the Poisson equation (2) with a constant $\nabla\phi$ and constant nonzero μ on \mathbb{R}^4 . Thus, in Newtonian physics, cosmology never became a rational science.¹ General relativity, on the other hand, does admit homogeneous and isotropic solutions as well as their perturbations. Indeed, cosmological solutions of the Einstein equations were studied by Einstein himself, de Sitter, Friedmann, and Lemaitre in the early years of the subject.

When general relativity was formulated, the prevailing view was that the universe should be static. This led Einstein to add a term $\Lambda g_{\mu\nu}$ to the left-hand side of his equations, fine-tuned so as to allow for such a solution. The constant Λ is known as the *cosmological constant*. The expansion of the universe is now considered to be an observational fact, beginning with the fundamental discoveries of Hubble. Expanding universes can be modeled to a first approximation by so-called Friedmann-Lemaitre solutions to the Einstein-Euler system, with various values of Λ . In the past direction, these solutions are singular: this singular behavior is often given the suggestive name “the big bang.”

4.5 Future Developments

The plethora of exact solutions of the Einstein equations gives us a taste of what the qualitative behavior

1. One can study “Newtonian cosmology” by modifying the foundations of the Newtonian theory so as to describe the theory with a nonmetric connection on, say, $\mathbb{T}^3 \times \mathbb{R}$. But this step is of course inspired by general relativity (see section 3.5).

of more general solutions may be. But a true qualitative understanding of the nature of general solutions has been achieved only in a neighborhood of the very simplest solutions. The question of the stability of the black hole solutions described above remains unanswered, as do the cosmic censorship conjectures and the nature of the singularities that occur generically in general relativity. Yet these questions are fundamental to the physical interpretation of the theory, and indeed to assessing its very validity.

How likely is it that these questions can ever be answered by rigorous mathematics? Problems concerning the singular behavior of nonlinear hyperbolic partial differential equations are notoriously difficult. The rich geometric structure of the Einstein equations appears at first as a formidable additional complication, but it may also turn out to be a blessing. One can only hope that the Einstein equations will continue to reveal beautiful mathematical structure that answers fundamental questions about our physical world.

Further Reading

- Christodoulou, D. 1999. On the global initial value problem and the issue of singularities. *Classical Quantum Gravity* 16:A23–A35.
- Hawking, S. W., and G. F. R. Ellis. 1973. *The Large Scale Structure of Space-Time*. Cambridge Monographs on Mathematical Physics, number 1. Cambridge: Cambridge University Press.
- Penrose, R. 1965. Gravitational collapse and space-time singularities. *Physical Review Letters* 14:57–59.
- Rendall, A. 2008. *Partial Differential Equations in General Relativity*. Oxford: Oxford University Press.
- Weyl, H. 1919. *Raum, Zeit, Materie*. Berlin: Springer. (Also published in English, in 1952, as *Space, Time, Matter*. New York: Dover.)

IV.14 Dynamics

Bodil Branner

1 Introduction

Dynamical systems are used to describe the way systems evolve in time, and have their origin in the laws of nature that NEWTON [VI.14] formulated in *Principia Mathematica* (1687). The associated mathematical discipline, the theory of dynamics, is closely related to many parts of mathematics, in particular analysis, topology, measure theory, and combinatorics. It is also highly influenced and stimulated by problems

from the natural sciences, such as celestial mechanics, hydrodynamics, statistical mechanics, meteorology, and other parts of mathematical physics, as well as reaction chemistry, population dynamics, and economics.

Computer simulations and visualizations play an important role in the development of the theory; they have changed our views about what should be considered typical, rather than special and atypical.

There are two main branches of dynamical systems: continuous and discrete. The main focus of this paper will be *holomorphic dynamics*, which concerns discrete dynamical systems of a special kind. These systems are obtained by taking a HOLOMORPHIC FUNCTION [I.3 §5.6] f defined on the complex numbers and applying it repeatedly. An important example is when f is a quadratic polynomial.

1.1 Two Basic Examples

It is interesting to note that both types of dynamical system, continuous and discrete, can be well illustrated by examples that date back to Newton.

(i) *The N-body problem* models the motion in the solar system of the sun and $N - 1$ planets, and does so in terms of differential equations. Each body is represented by a single point, namely its center of mass, and the motion is determined by Newton's *universal law of gravitation*—also called the *inverse square law*. This says that the gravitational force between two bodies is proportional to each of their masses and inversely proportional to the square of the distance between them. Let \mathbf{r}_i denote the position vector of the i th body, m_i its mass, and g the universal gravitational constant. Then the force on the i th body due to the j th has magnitude $gm_i m_j / \|\mathbf{r}_j - \mathbf{r}_i\|^2$, and its direction is along the line from \mathbf{r}_i to \mathbf{r}_j . We can work out the total force on the i th body by adding up all these forces for $j \neq i$. Since a unit vector in the direction from \mathbf{r}_i to \mathbf{r}_j is $(\mathbf{r}_j - \mathbf{r}_i) / \|\mathbf{r}_j - \mathbf{r}_i\|$, we obtain a force of

$$g \sum_{j \neq i} m_i m_j \frac{\mathbf{r}_j - \mathbf{r}_i}{\|\mathbf{r}_j - \mathbf{r}_i\|^3}.$$

(There is a cube on the bottom rather than a square in order to compensate for the magnitude of $\mathbf{r}_j - \mathbf{r}_i$.) A solution to the N -body problem is a set of differentiable vector functions $(\mathbf{r}_1(t), \dots, \mathbf{r}_N(t))$, depending on time t , that satisfy the N differential equations

$$m_i \mathbf{r}_i''(t) = g \sum_{j \neq i} m_i m_j \frac{\mathbf{r}_j(t) - \mathbf{r}_i(t)}{\|\mathbf{r}_j(t) - \mathbf{r}_i(t)\|^3},$$

which result from Newton's second law, which states that force = mass \times acceleration.

Newton was able to solve the two-body problem explicitly. By neglecting the influence of other planets, he derived the laws formulated by Johannes Kepler, which describe how each planet moves in an elliptic orbit around the sun. However, the jump to $N > 2$ makes an enormous difference to the complication of the problem: except in very special cases, the system of equations can no longer be solved explicitly (see THE THREE-BODY PROBLEM [V.36]). Nevertheless, Newton's equations are of great practical importance when it comes to guiding satellites and other space missions.

(ii) NEWTON'S METHOD [II.4 §2.3] for solving equations is quite different and does not involve differential equations. We consider a differentiable function f of one real variable and wish to determine a zero of f , that is, a solution to the equation $f(x) = 0$. Newton's idea was to define a new function:

$$N_f(x) = x - \frac{f(x)}{f'(x)}.$$

To put this more geometrically, $N_f(x)$ is the x -coordinate of the point where the tangent line to the graph $y = f(x)$ at the point $(x, f(x))$ crosses the x -axis. (If $f'(x) = 0$, then this tangent line is horizontal and $N_f(x)$ is not defined.)

Under many circumstances, if x is close to a zero of f , then $N_f(x)$ is significantly closer. Therefore, if we start with some value x_0 and form the sequence obtained by repeated application of N_f , that is, the sequence x_0, x_1, x_2, \dots , where $x_1 = N_f(x_0)$, $x_2 = N_f(x_1)$, and so on, we can expect that this sequence will converge to a zero of f . And this is true: if the initial value x_0 is sufficiently close to a zero, then the sequence does indeed converge toward that zero, and does so extremely quickly, basically doubling the number of correct digits in each step. This rapid convergence makes Newton's method very useful for numerical computations.

1.2 Continuous Dynamical Systems

We can think of a *continuous dynamical system* as a system of first-order differential equations, which determine how the system evolves in time. A solution is called an *orbit* or *trajectory*, and is parametrized by a number t , which one usually thinks of as time, that takes real values and varies continuously: hence the name "continuous" dynamical system. A *periodic orbit*

of *period* T is a solution that repeats itself after time T , but not earlier.

The differential equation $x''(t) = -x(t)$ is of second order, but it is nevertheless a continuous dynamical system because it is equivalent to the system of two first-order differential equations $x'_1(t) = x_2(t)$ and $x'_2(t) = -x_1(t)$. In a similar way, the system of differential equations of the N -body problem can be brought into standard form by introducing new variables. The equations are equivalent to a system of $6N$ first-order differential equations in the variables of the position vectors $\mathbf{r}_i = (x_{i1}, x_{i2}, x_{i3})$ and the velocity vectors $\mathbf{r}'_i = (y_{i1}, y_{i2}, y_{i3})$. Thus, the N -body problem is a good example of a continuous dynamical system.

In general, if we have a dynamical system consisting of n equations, then we can write the i th equation in the form

$$x'_i(t) = f_i(x_1(t), \dots, x_n(t)),$$

or alternatively we can write all the equations at once in the form $\mathbf{x}'(t) = \mathbf{f}(\mathbf{x}(t))$, where $\mathbf{x}(t)$ is the vector $(x_1(t), \dots, x_n(t))$ and $\mathbf{f} = (f_1, \dots, f_n)$ is a function from \mathbb{R}^n to \mathbb{R}^n . Note that \mathbf{f} is assumed not to depend on t . If it does, then the system can be brought into standard form by adding the variable $x_{n+1} = t$ and the differential equation $x'_{n+1}(t) = 1$, which increases the dimension of the system from n to $n + 1$.

The simplest systems are *linear* ones, where \mathbf{f} is a linear map: that is, $\mathbf{f}(\mathbf{x})$ is given by $A\mathbf{x}$ for some constant $n \times n$ matrix A . The system above, $x'_1(t) = x_2(t)$ and $x'_2(t) = -x_1(t)$, is an example of a linear system. Most systems, however, including the one for the N -body problem, are *nonlinear*. If the function \mathbf{f} is "nice" (for instance, differentiable), then *uniqueness* and *existence* of solutions are guaranteed for any initial point \mathbf{x}_0 . That is, there is exactly one solution that passes through the point \mathbf{x}_0 at time $t = 0$. For example, in the N -body problem there is exactly one solution for any given set of initial position vectors and initial velocity vectors. It also follows from uniqueness that any pair of orbits must either coincide or be totally disjoint. (Bear in mind that the word "orbit" in this context does not mean the set of positions of a single point mass, but rather the evolution of the vector that represents all the positions and velocities of all the masses.)

Although it is seldom possible to express solutions to nonlinear systems explicitly, we know that they exist, and we call the dynamical system *deterministic* since solutions are completely determined by their initial conditions. For a given system and given initial con-

ditions it is therefore theoretically possible to predict its entire future evolution.

1.3 Discrete Dynamical Systems

A *discrete dynamical system* is a system that evolves in jumps: “time,” in such a system, is best represented by an integer rather than a real number. A good example is Newton’s method for solving equations. In this instance, the sequence of points we saw earlier, $x_0, x_1, \dots, x_k, \dots$, where $x_k = N_f(x_{k-1})$, is called the *orbit* of x_0 . We say that it is obtained by *iteration* of the function N_f , i.e., by repeated application of the function.

This idea can easily be generalized to other mappings $F : X \rightarrow X$, where X could be the real axis, an interval in the real axis, the plane, a subset of the plane, or some more complicated space. The important thing is that the output $F(x)$ of any input x can be used as the next input. This guarantees that the orbit of any x_0 in X is defined for all future times. That is, we can define a sequence, $x_0, x_1, \dots, x_k, \dots$, where $x_k = F(x_{k-1})$ for every k . If the function F has an inverse F^{-1} , then we can iterate both forwards and backwards and obtain the *full orbit* of x_0 as the bi-infinite sequence $\dots, x_{-2}, x_{-1}, x_0, x_1, x_2, \dots$, where $x_k = F(x_{k-1})$ and, equivalently, $x_{k-1} = F^{-1}(x_k)$, for all integer values.

The orbit of x_0 is *periodic* of period k if it repeats itself after time k , but not earlier, i.e., if $x_k = x_0$, but $x_j \neq x_0$ for $j = 1, \dots, k-1$. The orbit is called *pre-periodic* if it is eventually periodic, in other words if there exist $\ell \geq 1$ and $k \geq 1$ such that x_ℓ is periodic of period k , but none of the x_j for $0 \leq j < \ell$ are periodic. The notion of pre-periodicity has no counterpart in continuous dynamics.

A discrete dynamical system is deterministic, since the orbit of any given initial point x_0 is completely determined once you know x_0 .

1.4 Stability

The modern theory of dynamics was greatly influenced by the work of POINCARÉ [VI.61], and in particular by his prize-winning memoir on the 3-body problem, succeeded by three more elaborate volumes on celestial mechanics, all from the late nineteenth century. The memoir was written in response to a competition where one of the proposed problems concerned stability of the solar system. Poincaré introduced the so-called *restricted 3-body problem*, where the third body

is assumed to have an infinitely small mass: it does not influence the motion of the other two bodies but it is influenced by them. Poincaré’s work became the prelude to *topological dynamics*, which focuses on topological properties of solutions to dynamical systems and takes a qualitative approach to them.

Of special interest is the long-term behavior of a system. A periodic orbit is called *stable* if all orbits through points sufficiently close to it stay close to it at all future times. It is called *asymptotically stable* if all sufficiently close orbits approach it as time tends to infinity. Let us illustrate this by two linear examples in discrete dynamics. For the real function $F(x) = -x$, all points have a periodic orbit: 0 has period 1 and all other x have period 2. Every orbit is stable, but none is asymptotically stable. The real function $G(x) = \frac{1}{2}x$ has only one periodic orbit, namely 0. Since $G(0) = 0$, this orbit has period 1, and we call it a *fixed point*. If you take any number and repeatedly divide it 2, then the resulting sequence will approach 0, so the fixed point 0 is asymptotically stable.

One of the methods introduced by Poincaré during his study of the 3-body problem was a reduction from a continuous dynamical system, in dimension n , say, to an associated discrete dynamical system, a mapping in dimension $n-1$. The idea is as follows. Suppose we have a periodic orbit of period $T > 0$ in some continuous system. Choose a point \mathbf{x}_0 on the orbit and a hypersurface Σ through \mathbf{x}_0 , for instance part of a hyperplane, such that the orbit cuts through Σ at \mathbf{x}_0 . For any point in Σ that is sufficiently close to \mathbf{x}_0 , one can follow its orbit around and see where it next intersects Σ . This defines a transformation, known as the *Poincaré map*, which takes the original point to the next point of intersection of its orbit with Σ . It follows from the fact that dynamical systems have unique solutions that every Poincaré map is injective in the neighborhood of \mathbf{x}_0 (within Σ) for which the Poincaré map is defined. One can perform both forwards and backwards iterations. Note that the periodic orbit of \mathbf{x}_0 in the continuous system is stable (respectively, asymptotically stable) exactly when the fixed point \mathbf{x}_0 of the Poincaré map in the discrete system is stable (respectively, asymptotically stable).

1.5 Chaotic Behavior

The notion of *chaotic dynamics* arose in the 1970s. It has been used in different settings, and there is no single definition that covers all uses of the term. However, the property that best characterizes chaos is the phenomenon of *sensitive dependence on initial conditions*.

Poincaré was the first to observe sensitivity to initial conditions in his treatment of the 3-body problem.

Instead of describing his observations let us look at a much simpler example from discrete dynamics. Take as a dynamical space X the half-open unit interval $[0, 1)$, and let F be the function that doubles a number and reduces it modulo 1. That is, $F(x) = 2x$ when $0 \leq x < \frac{1}{2}$ and $F(x) = 2x - 1$ when $\frac{1}{2} \leq x < 1$. Let x_0 be a number in X and let its iterates be $x_1 = F(x_0)$, $x_2 = F(x_1)$, and so on. Then x_k is the fractional part of $2^k x_0$. (The fractional part of a real number t is what you get when you subtract the largest integer less than t .)

A good way to understand the behavior of the sequence x_0, x_1, x_2, \dots of iterates is to consider the binary expansion of x_0 . Suppose, for example, that this begins $0.110100010100111\dots$. To double a number when it is written in binary, all you have to do is shift every digit to the left (just as one does in the decimal system when multiplying by 10). So $2x_0$ will have a binary expansion that begins $1.10100010100111\dots$. To obtain $F(x_0)$, we have to take the fractional part of this, which we do by subtracting the initial 1. This gives us $x_1 = 0.10100010100111\dots$. Repeating the process we find that $x_2 = 0.0100010100111\dots$, $x_3 = 0.100010100111\dots$, and so on. (Notice that when we calculated x_3 from x_2 there was no need to subtract 1, since the first digit after the “decimal point” was a 0.) Now consider a different choice of initial number, $x'_0 = 0.110100010110110\dots$. The first nine digits after the decimal point are the same as the first nine digits of x_0 , so x'_0 is very close to x_0 . However, if we apply F nine times to x_0 and x'_0 , then their respective tenth digits have shifted leftwards and become the first digits of $x_9 = 0.00111\dots$ and $x'_9 = 0.10110\dots$. These two numbers differ by almost $\frac{1}{2}$, so they are not at all close.

In general, if we know x_0 to an accuracy of k binary digits and no more, then after k iterations of the map F we have lost all information: x_k could lie anywhere in the interval $[0, 1)$. Therefore, even though the system is deterministic, it is impossible to predict its long-term behavior without knowing x_0 with perfect accuracy.

This is true in general: it is impossible to make long-term predictions in any part of a dynamical system that shows sensitivity to initial conditions unless the initial conditions are known exactly. In practical applications this is never the case. For instance, when applying a mathematical model to perform weather forecasts, one does not know the initial conditions exactly, and this is why reliable long-term forecasting is impossible.

Sensitivity is also important in the notion of so-called *strange attractors*. A set A is called an *attractor* if all orbits that start in A stay in A and if all orbits through nearby points get closer and closer to A . In continuous systems, some simple sets that can be attractors are equilibrium points, periodic orbits (limit cycles), and surfaces such as a torus. In contrast to these examples, strange attractors have both complicated geometry and complicated dynamics: the geometry is *fractal* and the dynamics sensitive. We shall see examples of fractals later on.

The best-known strange attractor is the *Lorenz attractor*. In the early 1960s, the meteorologist Edward N. Lorenz studied a three-dimensional continuous dynamical system that gave a simplified model of heat flow. While doing so, he noticed that if he restarted his computer with its initial conditions chosen as the output of an earlier calculation, then the trajectory started to diverge from the one he had previously observed. The explanation he found was that the computer used more precision in its internal calculations than it showed in its output. For this reason, it was not immediately apparent that the initial conditions were in fact very slightly different from before. Because the system was sensitive, this tiny difference eventually made a much bigger difference. He coined the poetic phrase “the butterfly effect” to describe this phenomenon, suggesting that a small disturbance such as a butterfly flickering its wings could in time have a dramatic effect on the long-term evolution of the weather and trigger a tornado thousands of miles away. Computer simulations of the Lorenz system indicate that solutions are attracted to a complicated set that “looks like” a strange attractor. The question of whether it actually was one remained open for a long time. It is not obvious how trustworthy computer simulations are when one is studying sensitive systems, since the computer rounds off the numbers in each step. In 1998 Warwick Tucker gave a computer-assisted proof that the Lorenz attractor is in fact a strange attractor. He used *interval arithmetic*, where numbers are represented by intervals and estimates can be made precise.

For topological reasons, sensitivity to initial conditions is possible for continuous dynamical systems only when the dimension is at least 3. For discrete systems where the map F is injective, the dimension must be at least 2. However, for noninjective mappings, sensitivity can occur for one-dimensional systems, as we saw with the example given earlier. This is one of

the reasons that discrete one-dimensional dynamical systems have been intensively studied.

1.6 Structural Stability

Two dynamical systems are said to be *topologically equivalent* if there is a homeomorphism (a continuous map with continuous inverse) that maps the orbits of one system onto the orbits of the other, and vice versa. Roughly speaking, this means that there is a continuous change of variables that turns one system into the other.

As an example, consider the discrete dynamical system given by the real quadratic polynomial $F(x) = 4x(1 - x)$. Suppose we were to make the substitution $y = -4x + 2$. How could we describe the system in terms of y ? Well, if we apply F , then we change x to $4x(1 - x)$, which means that $y = -4x + 2$ changes $F(x)$ to $-4F(x) + 2 = -16x(1 - x) + 2$. But

$$\begin{aligned} -16x(1 - x) + 2 &= 16x^2 - 16x + 2 \\ &= (-4x + 2)^2 - 2 \\ &= y^2 - 2. \end{aligned}$$

Therefore, the effect of applying the polynomial function F to x is to apply a different polynomial function to y , namely $Q(y) = y^2 - 2$. Since the change of variables from x to $-4x + 2$ is continuous and invertible, one says that the functions F and Q are *conjugate*.

Because F and Q are conjugate, the orbit of any x_0 under F becomes, after the change of variables, the orbit of the corresponding point $y_0 = -4x_0 + 2$ under Q . That is, for every k we have $y_k = -4x_k + 2$. The two systems are topologically equivalent: if you want to understand the dynamics of one of them, you can if you study the other, since its dynamics will be qualitatively the same.

For continuous dynamical systems the notion of equivalence is slightly looser in that we allow a homeomorphism between two topologically equivalent systems to map one orbit onto another without respecting the exact time evolution, but for discrete dynamical systems we must demand that the time evolution is respected as in the example above: in other words, we insist on conjugacy.

The term *dynamical system* was coined by Stephen Smale in the 1960s and has taken off since then. Smale evolved the theory of *robust* systems, also named *structurally stable* systems, a notion that was introduced in the 1930s by Alexander A. Andronov and Lev S. Pontryagin. A dynamical system is called structurally stable if all systems sufficiently close to it, belonging to

some specified family of systems, are in fact topologically equivalent to it. We say that they all have the same qualitative behavior. An example of the kind of family one might consider is the set of all real quadratic polynomials of the form $x^2 + a$. This family is parametrized by a , and the systems close to a given polynomial $x^2 + a_0$ are all the polynomials $x^2 + a$ for which a is close to a_0 . We shall return to the question of structural stability when we discuss holomorphic dynamics later.

If a family of dynamical systems parametrized by a variable a is not structurally stable, it may still be that the system with parameter a_0 is topologically equivalent to all systems with parameter a in some region that contains a_0 . A major goal of research into dynamics is to understand not just the qualitative structure of each system in the family, but also the structure of the *parameter space*, that is, how it is divided up into such regions of stability. The boundaries that separate these regions form what is called the *bifurcation set*: if a_0 belongs to this set, then there will be parameters a arbitrarily close to a_0 for which the corresponding system has a different qualitative behavior.

A description and classification of structurally stable systems and a classification of possible bifurcations is not within reach for general dynamical systems. However, one of the success stories in the subject, holomorphic dynamics, studies a special class of dynamical systems for which many of these goals have been attained. It is time to turn our attention to this class.

2 Holomorphic Dynamics

Holomorphic dynamics is the study of discrete dynamical systems where the map to be iterated is a HOLOMORPHIC FUNCTION [I.3 §5.6] of the COMPLEX NUMBERS [I.3 §1.5]. Complex numbers are typically denoted by z . In this article, we shall consider iterations of complex polynomials and rational functions (that is, functions like $(z^2 + 1)/(z^3 + 1)$ that are ratios of polynomials), but much of what we shall say about them is true for more general holomorphic functions, such as EXPONENTIAL [III.25] and TRIGONOMETRIC [III.94] functions.

Whenever one restricts attention to a special kind of dynamical system, there will be tools that are specially adapted to that situation. In holomorphic dynamics these tools come from complex analysis. When we concentrate on rational functions, there are more special tools, and if we restrict further to polynomials, then there are yet others, as we shall see.

Why might one be interested in iterating rational functions? One answer arose in 1879, when CAYLEY [VI.46] had the idea of trying to find roots of complex polynomials by extending Newton's method, which we discussed in the introduction, from real numbers to complex numbers. Given any polynomial P , the corresponding Newton function N_P is a rational function, given by the formula

$$N_P(z) = z - \frac{P(z)}{P'(z)} = \frac{zP'(z) - P(z)}{P'(z)}.$$

To apply Newton's method, one iterates this rational function.

The study of the iteration of rational functions flourished at the beginning of the twentieth century, thanks in particular to work of Pierre Fatou and Gaston Julia (who independently obtained many of the same results). Part of their work concerned the study of the local behavior of functions in the neighborhoods of a fixed point. But they were also concerned about global dynamical properties and were inspired by the theory of so-called *normal families*, then recently established by Paul Montel. However, research on holomorphic dynamics almost came to a stop around 1930, because the fractal sets that lay behind the results were so complicated as to be almost beyond imagination. The research came back to life in around 1980 with the vastly extended calculating powers of computers, and in particular the possibility of making sophisticated graphic visualizations of these fractal sets. Since then, holomorphic dynamics has attracted a lot of attention. New techniques continue to be developed and introduced.

To set the scene, let us start by looking at one of the simplest of polynomials, namely z^2 .

2.1 The Quadratic Polynomial z^2

The dynamics of the simplest quadratic polynomial, $Q_0(z) = z^2$, plays a fundamental role in the understanding of the dynamics of any quadratic polynomial. Moreover, the dynamical behavior of Q_0 can be analyzed and understood completely.

If $z = re^{i\theta}$, then $z^2 = r^2e^{2i\theta}$, so squaring a complex number squares its modulus and doubles its argument. Therefore, the unit circle (the set of complex numbers of modulus 1) is mapped by Q_0 to itself, while a circle of radius $r < 1$ is mapped onto a circle closer to the origin, and a circle of radius $r > 1$ is mapped onto a circle farther away.

Let us look more closely at what happens to the unit circle. A typical point in the circle, $e^{i\theta}$, can be parametrized by its argument θ , which we can take to lie in the interval $[0, 2\pi)$. When we square this number, we obtain $e^{2i\theta}$, which is parametrized by the number 2θ if $2\theta < 2\pi$, but if $2\theta \geq 2\pi$, then we subtract 2π so that the argument, $2\theta - 2\pi$, still lies in $[0, 2\pi)$. This is strongly reminiscent of the dynamical system we considered in section 1.5. In fact, if we replace the argument θ by its *modified argument* $\theta/2\pi$, which amounts to writing $e^{2\pi i\theta}$ instead of $e^{i\theta}$, then it becomes exactly the same system. Therefore, the behavior of z^2 on the unit circle is chaotic.

As for the rest of the complex plane, the origin is an asymptotically stable fixed point, $Q_0(0) = 0$. For any point z_0 inside the unit circle the iterates z_k converge to 0 as k tends to infinity. For any point z_0 outside the unit circle the distance $|z_k|$ between the iterates z_k and the origin tends to infinity as k tends to infinity. The set of initial points z_0 with bounded orbit is equal to the closed unit disk, i.e., all points for which $|z_0| \leq 1$. Its boundary, the unit circle, divides the complex plane into two domains with qualitatively different dynamical behavior.

Some orbits of Q_0 are periodic. In order to determine which ones, we first notice that the only possibility outside the unit circle is the fixed point at the origin, since all other points, when you repeatedly square them, either get steadily closer and closer to the origin, or get steadily farther and farther away. So now let us look at the unit circle, and consider the point $e^{2\pi i\theta_0}$, with modified argument θ_0 . If this point is periodic with period k , we must have $2^k\theta_0 = \theta_0 \pmod{1}$: that is, $(2^k - 1)\theta_0$ must be an integer. Because of this, it is convenient to parametrize a point on the unit circle by its modified argument. From now on, when we say "the point θ ," we shall mean the point $e^{2\pi i\theta}$, and when we say "argument" we shall mean modified argument.

We have just established that the point θ is periodic with period k only if $(2^k - 1)\theta$ is an integer. It follows that there is one point of period 1, namely $\theta_0 = 0$. There are two points of period 2, forming one orbit, namely $\frac{1}{3} \mapsto \frac{2}{3} \mapsto \frac{1}{3}$. There are six points for period 3, forming two orbits, namely $\frac{1}{7} \mapsto \frac{2}{7} \mapsto \frac{4}{7} \mapsto \frac{1}{7}$ and $\frac{3}{7} \mapsto \frac{6}{7} \mapsto \frac{5}{7} \mapsto \frac{3}{7}$. (At each stage, we double the number we have, and subtract 1 if that is needed to get us back into the interval $[0, 1)$.) The points of period 4 are fractions with denominator 15, but the converse is not true: the fractions $\frac{3}{15} = \frac{1}{5}$ and $\frac{6}{15} = \frac{2}{5}$ have the

lower period 2. The periodic points on the unit circle are *dense* in the unit circle, meaning that arbitrarily close to any point is a periodic point. This follows from the observation that all repeating binary expansions, such as $0.1100011000110001100011000\dots$ are periodic, and any finite sequence of 0s and 1s is the start of a repeating sequence. One can, in fact, show that the periodic points on the unit circle are exactly the points whose argument is a fraction p/q in $[0, 1)$ with q odd. Any fraction with even denominator can be written in the form $p/(2^\ell q)$ for some odd number q . After ℓ iterations, such a fraction will land on a periodic point, so the initial point is pre-periodic. Points with rational argument in $[0, 1)$ have a finite orbit, while points with irrational argument have an infinite orbit. The reason for taking modified arguments is now justified: the behavior of the dynamics depends on whether θ_0 is rational or irrational.

When θ_0 is irrational its orbit may or may not be dense in $[0, 1)$. This is another fact that is easy to see if one considers binary expansions. For instance, a very special example of a θ_0 with a dense orbit is given by the binary expansion

$$\theta_0 = 0.0100011011000001010011100101110111\dots,$$

where one obtains this expansion by simply listing all finite binary sequences in turn: first the blocks of length one, 0 and 1, then the blocks of length two, 00, 01, 10, and 11, and so on. When we iterate, this binary expansion shifts to the left and all possible finite sequences appear at some time or another at the beginning of some iterate θ_k .

2.2 Characterization of Periodic Points

Let z_0 be a fixed point of a holomorphic map F . How do the iterates of points near z_0 behave? The answer depends crucially on a number ρ , called the *multiplier* of the fixed point, which is defined to be $F'(z_0)$. To see why this is relevant, notice that if z is very close to z_0 , then $F(z)$ is, to a first-order approximation, equal to $F(z_0) + F'(z_0)(z - z_0) = z_0 + \rho(z - z_0)$. Thus, when you apply F to a point near z_0 , its difference from z_0 approximately multiplies by ρ . If $|\rho| < 1$, then nearby points will get closer to z_0 , in which case z_0 is called an *attracting* fixed point. If $\rho = 0$, then this happens very quickly and z_0 is called *super-attracting*. If $|\rho| > 1$, then nearby points get farther away and z_0 is called *repelling*. Finally, if $|\rho| = 1$, then one says that z_0 is *indifferent*.

If z_0 is indifferent, then its multiplier will take the form $\rho = e^{2\pi i\theta}$, and near z_0 the map F will be approximately a rotation about z_0 by an angle of $2\pi\theta$. The behavior of the system depends very much on the precise value of θ . We call the fixed point *rationally* or *irrationally indifferent* if θ is rational or irrational, respectively. The dynamics is not yet completely understood in all irrational cases.

A periodic point z_0 of period k will be a fixed point of the k th iterate $F^k = F \circ \dots \circ F$ of F . For this reason we define its multiplier by $\rho = (F^k)'(z_0)$. It follows from the chain rule that

$$(F^k)'(z_0) = \prod_{j=0}^{k-1} F'(z_j)$$

and therefore that the derivative of F^k is the same at all points of the periodic orbit. This formula also implies that a super-attracting periodic orbit must contain a critical point (that is, a point where the derivative of F is zero): if $(F^k)'(z_0) = 0$, then at least one $F'(z_j)$ must be 0.

Note that 0 is a super-attracting fixed point of Q_0 , and that any periodic orbit of Q_0 of period k on the unit circle has multiplier 2^k . All periodic orbits on the unit circle are therefore repelling.

2.3 A One-Parameter Family of Quadratic Polynomials

The quadratic polynomial Q_0 sits at the center of the one-parameter family of quadratic polynomials of the form $Q_c(z) = z^2 + c$. (We considered this family earlier, but then z and c were real rather than complex.) For each fixed complex number c we are interested in the dynamics of the polynomial Q_c under iteration. The reason we do not need to study more general quadratic polynomials is that they can be brought into this form by a simple substitution $w = az + b$, similar to the substitution in the real example in section 1.6. For any given quadratic polynomial P we can find exactly one substitution $w = az + b$ and one c such that

$$a(P(z)) + b = (az + b)^2 + c \quad \text{for all } z.$$

Therefore, if we understand the dynamics of the polynomials Q_c , then we understand the dynamics of all quadratic polynomials.

There are other representative families of quadratic polynomials that can be useful. One example is the family $F_\lambda(z) = \lambda z + z^2$. The substitution $w = z + \frac{1}{2}\lambda$ changes F_λ into Q_c , where $c = \frac{1}{2}\lambda - \frac{1}{4}\lambda^2$. We shall return to the expression of c in terms of λ later on. In

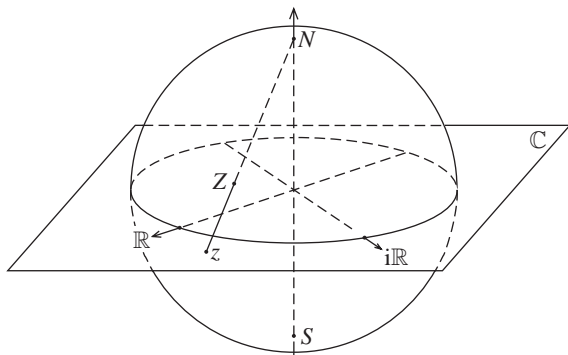


Figure 1 The Riemann sphere.

the family of polynomials Q_c , the parameter $c = Q_c(0)$ coincides with the only *critical value* of Q_c in the plane: as we shall see later, critical orbits play an essential role in the analysis of the global dynamics. In the family of polynomials F_λ the parameter λ is equal to the multiplier of the fixed point at the origin of F_λ , which sometimes makes this family more convenient.

2.4 The Riemann Sphere

To understand further the dynamics of polynomials it is best to regard them as a special case of rational functions. Since a rational function can sometimes be infinite, the natural space to consider is not the complex plane \mathbb{C} but the *extended complex plane*, which is the complex plane together with the point “ ∞ .” This space is denoted $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$. A geometrical picture (see figure 1) is obtained by identifying the extended complex plane with the *Riemann sphere*. This is simply the unit sphere $\{(x_1, x_2, x_3) : x_1^2 + x_2^2 + x_3^2 = 1\}$ in three-dimensional space. Given a number z in the complex plane, the straight line joining z to the north pole $N = (0, 0, 1)$ intersects this sphere in exactly one place (apart from N itself). This place is the point in the sphere that is associated with z . Notice that the bigger $|z|$ is, the closer the associated point is to N . We therefore regard N as corresponding to the point ∞ .

Let us now think of $Q_0(z) = z^2$ as a function from $\hat{\mathbb{C}}$ to $\hat{\mathbb{C}}$. We have seen that 0 is a super-attracting fixed point of Q_0 . What about ∞ , which is a fixed point as well? The classification we gave in terms of multipliers does not work at ∞ , but a standard trick in this situation is to “move” ∞ to 0. If one wishes to understand the behavior of a function f with a fixed point at ∞ , one can look instead at the function $g(z) = 1/f(1/z)$, which has a fixed point at 0 (since $1/f(1/0) = 1/f(\infty) =$

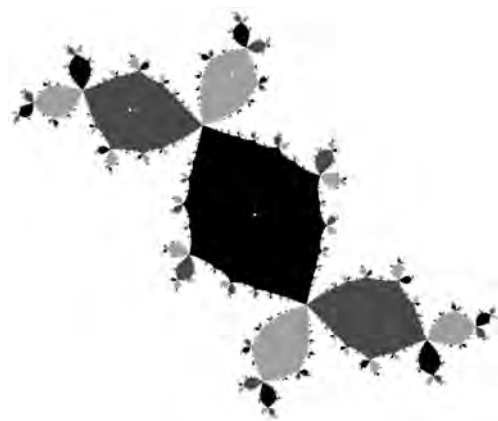


Figure 2 The *Douady rabbit*. The filled Julia set of Q_{c_0} where c_0 is the one root of the polynomial $(c^2 + c)^2 + c$ that has positive imaginary part. This corresponds to one of the three possible c values for which the critical orbit $0 \mapsto c \mapsto c^2 + c \mapsto (c^2 + c)^2 + c = 0$ is periodic of period 3. The critical orbit is marked with three white dots inside the filled Julia set: 0 in the black, c_0 in the light gray, and $c_0^2 + c_0$ in the gray. The corresponding three attracting basins of $Q_{c_0}^3$ are marked in black, light gray, and gray, respectively. The Julia set is the common boundary of the black, light gray, and gray basins of attraction as well as of $A_{c_0}(\infty)$.

$1/\infty = 0$). When $f(z) = z^2$, $g(z)$ is also z^2 , so ∞ is also a super-attracting fixed point of Q_0 .

In general, if P is any nonconstant polynomial, then it is natural to define $P(\infty)$ to be ∞ . Applying the above trick, we obtain a rational function. For example, if $P(z) = z^2 + 1$, then $1/P(1/z) = z^2/(z^2 + 1)$. If P has degree at least 2, then ∞ is a super-attracting fixed point.

The connection between $\hat{\mathbb{C}}$ and rational functions is expressed by the following fact: a function $F : \hat{\mathbb{C}} \rightarrow \hat{\mathbb{C}}$ is holomorphic everywhere (with suitable definitions at ∞) if and only if it is a rational function. This is not obvious, but is typically proved in a first course in complex analysis. Among the rational functions, the polynomials are the ones for which $F(\infty) = \infty = F^{-1}(\infty)$.

A polynomial P of degree d has $d - 1$ critical points in the plane (not including ∞). These are the roots of the derivative P' , counted with multiplicity. The critical point at ∞ has multiplicity $d - 1$, as can again be seen by looking at the map $1/P(1/z)$. In particular, quadratic polynomials have exactly one critical point in the plane. The degree of a rational function P/Q (where P and Q have no common roots) is defined to be the maximal degree of the polynomials P and Q . A rational function

of degree d has $2d - 2$ critical points in $\hat{\mathbb{C}}$, as we have just seen for polynomials.

2.5 Julia Sets of Polynomials

It can be shown that the only invertible holomorphic maps from \mathbb{C} to \mathbb{C} are polynomials of degree 1, that is, functions of the form $az + b$ with $a \neq 0$. The dynamical behavior of these maps is easy to analyze, simple, and hence not interesting.

From now on, therefore, we shall consider only polynomials P of degree at least 2. For all such polynomials, ∞ is a super-attracting fixed point, from which it follows that the plane is split into two disjoint sets with qualitatively different dynamics, one consisting of points that are attracted to ∞ and the other consisting of points that are not. The *attracting basin* of ∞ , denoted by $A_P(\infty)$, consists of all initial points z such that $P^k(z) \rightarrow \infty$ as $k \rightarrow \infty$. (Here, $P^k(z)$ stands for the result of applying P to z k times.) The complement of $A_P(\infty)$ is called the *filled Julia set*, and is denoted by K_P . It can be defined as the set of all points z such that the sequence $z, P(z), P^2(z), P^3(z), \dots$ is bounded. (It is not hard to show that sequences of this kind either tend to ∞ or are bounded.)

The attracting basin of ∞ is an open set and the filled Julia set is a closed, bounded set (i.e., a COMPACT SET [III.9]). The attracting basin of ∞ is always connected. For this reason the boundary of K_P is equal to the boundary of $A_P(\infty)$. The common boundary is called the *Julia set* of P and is denoted by J_P . The three sets K_P , $A_P(\infty)$, and J_P are completely invariant, i.e., $P(K_P) = K_P = P^{-1}(K_P)$, and so on. If we replace P by any iterate P^k , then the filled Julia set, the attracting basin of ∞ , and the Julia set of P^k are the same sets as those of P .

For the polynomial Q_0 , we showed earlier that the filled Julia set is the closed unit disk, $\{z : |z| \leq 1\}$; the attracting basin of ∞ is its complement, $\{z : |z| > 1\}$; and the Julia set is the unit circle, $\{z : |z| = 1\}$.

The name “filled Julia set” refers to the fact that K_P is equal to J_P with all its holes (or, more formally, the bounded components of its complement) filled in. The complement of the Julia set is called the *Fatou set* and any connected component of it is called a *Fatou component*.

Figures 2–6 show different examples of Julia sets of quadratic polynomials Q_c . For simplicity we set $K_{Q_c} = K_c$, $A_{Q_c}(\infty) = A_c(\infty)$, and $J_{Q_c} = J_c$. Note that all

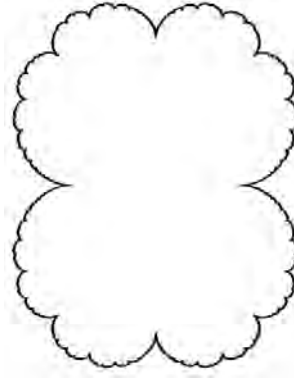


Figure 3 The Julia set of $Q_{1/4}$. Every point inside the Julia set (including the critical point 0) is attracted (under repeated applications of $Q_{1/4}$) to the rationally indifferent fixed point $\frac{1}{2}$ with multiplier $\rho = 1$, which belongs to $J_{1/4}$.

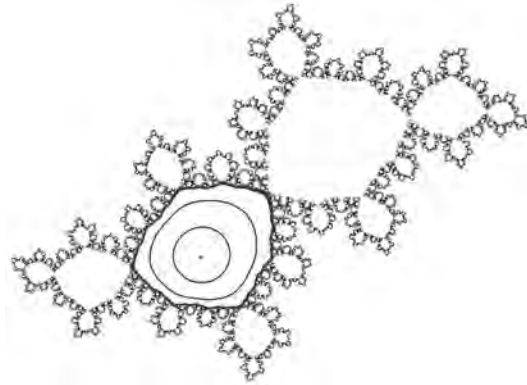


Figure 4 The Julia set of Q_c with a so-called *Siegel disk* around an irrationally indifferent fixed point of multiplier $\rho = e^{2\pi i(\sqrt{5}-1)/2}$. The corresponding c -value is equal to $\frac{1}{2}\rho - \frac{1}{4}\rho^2$. In the Siegel disk, the Fatou component containing the fixed point, the action of Q_c can, after a suitable change of variables, be expressed as $w \mapsto \rho w$. The fixed point is marked and so are some orbits of points in its vicinity. The critical orbit is dense in the boundary of the Siegel disk.

Julia sets J_c are symmetric around 0, owing to the symmetry in the formula: $Q_c(-z) = Q_c(z)$, which implies that if a point z belongs to J_c , then so does $-z$.

2.6 Properties of Julia Sets

In this section we shall list several common properties of Julia sets. The proofs of these, which are beyond the scope of this article, mostly depend on the theory of *normal families*.

Note to PUP: I still need to send Dimitri the figures from this article (or from The Companion as a whole) to see if any aren't up to scratch and need PUP redrawing. I will do that this week.

- The Julia set is the set of points for which the system displays sensitivity to initial conditions, i.e., the chaotic subset of the dynamical system.
- The repelling orbits belong to the Julia set and form a dense subset of the set. That is, any point in the Julia set can be approximated arbitrarily well by a repelling point. This is the definition originally used by Julia. (Of course, the name “Julia set” was used only later.)
- For any point z in the Julia set, the set of iterated preimages $\bigcup_{k=1}^{\infty} F^{-k}(z)$ forms a dense subset of the Julia set. This property is used when one is making computer pictures of Julia sets.
- In fact, for any point z in $\hat{\mathbb{C}}$ (with at most one or two exceptions), the closure of the set of iterated preimages contains the Julia set.
- For any point z in the Julia set and any neighborhood U_z of z , the iterated images $F^k(U_z)$ cover all of $\hat{\mathbb{C}}$ except at most one or two exceptional points. This property demonstrates an extreme sensitivity to initial conditions.
- If Ω is a union of Fatou components that is completely invariant (that is, $F(\Omega) = \Omega = F^{-1}(\Omega)$), then the boundary of Ω coincides with the Julia set. This justifies the definition of the Julia set of a polynomial as the boundary of the attracting basin of ∞ . Compare also with figure 2, where the attracting basins of $Q_{c_0}^3$ and $A_{c_0}(\infty)$ are examples of such completely invariant sets.
- The Julia set is either connected or consists of uncountably many connected components. An example of the latter is shown in figure 6.
- The Julia set is typically a fractal: when one zooms in on it, one finds that the complication of the set is repeated at all scales. It is also *self-similar*, in the following sense: for any noncritical point z in the Julia set, any sufficiently small neighborhood U_z of z is mapped bijectively onto $F(U_z)$, a neighborhood of $F(z)$. The Julia set in U_z and the Julia set in $F(U_z)$ look alike.

All but the last two properties can easily be verified in the example Q_0 . In this case the exceptional points are 0 and ∞ .

2.7 Böttcher Maps and Potentials

2.7.1 Böttcher Maps

Consider the quadratic polynomial $Q_{-2}(z) = z^2 - 2$. If z belongs to the interval $[-2, 2]$, then z^2 belongs to

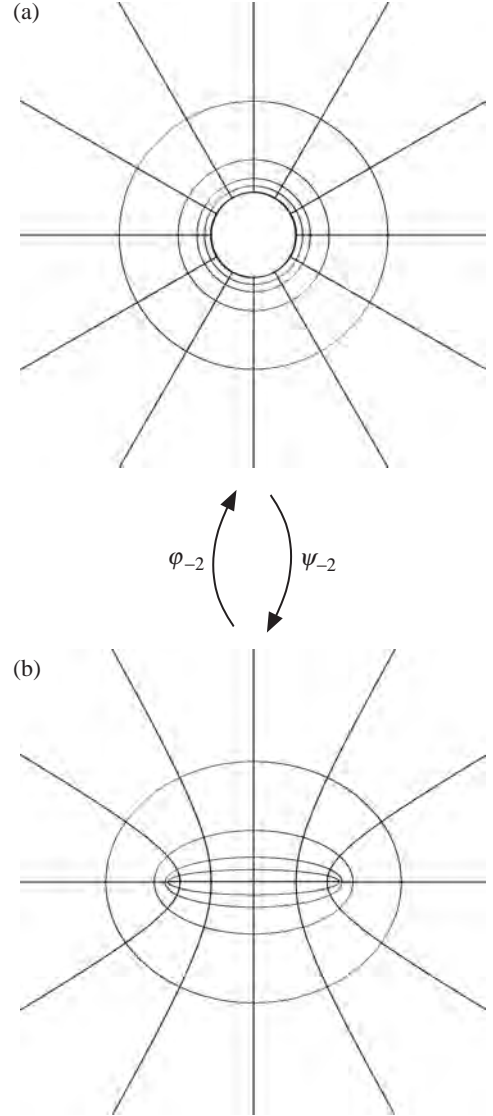


Figure 5 (a) Some equipotentials and external rays $\mathcal{R}_0(\theta)$ of Q_0 in $A_0(\infty)$, the set of complex numbers of modulus greater than 1. (b) The corresponding equipotentials and external rays $\mathcal{R}_{-2}(\theta)$ of Q_{-2} in $A_{-2}(\infty)$, the set of complex numbers not in $K_{-2} = J_{-2} = [-2, 2]$. The external rays that are drawn have arguments $\theta = \frac{1}{12}p$, where $p = 0, 1, \dots, 11$.

the interval $[0, 4]$, so $Q_{-2}(z)$ also belongs to the interval $[-2, 2]$. It follows that this interval is contained in the filled Julia set K_{-2} .

The polynomial $Q_{-2}(z)$ is not topologically equivalent to $Q_0(w) = w^2$, but when z is big enough, it behaves in a similar way, since 2 is small compared